

VANDERBILT UNIVERSITY



NASHVILLE, TENNESSEE 37203

TELEPHONE (615) 322-7311

Center for Mental
Health Policy

Leonard Bickman, Ph.D.
Director

Institute for Public Policy Studies • Box 7701, Station B • Direct phone 322-8435
Bitnet: Bickman@VUCTRVAX • FAX 322-7049

AD-A271 717



Response to the Directorate of Health Care Studies and
Clinical Investigations Final Report (Revised) Assessing
Power Analysis Approaches for the
Fort Bragg Evaluation Project

DTIC
ELECTE
OCT2 1993
S A D

This document has been approved
for public release and sale; its
distribution is unlimited.

July 28, 1993

93-25665



109

93 10 22 08 0

Below is a summary of Vanderbilt's response, requested by Health Services Command (HSC), to the final report prepared by the Directorate of Health Care Studies and Clinical Investigations (HCSCI) in June 1993. The HCSCI report was filed in NTIS as Report CR 93-002.

According to its abstract, the purpose of the HCSCI-sponsored report was to present ". . . a statistical review regarding an extension of the Fort Bragg Evaluation Project by Vanderbilt University Center for Mental Health Policy. It contains an assessment of two data collection plans using power analysis. The Monte Carlo power analysis performed by Vanderbilt University is also evaluated." The abstract concludes, "Based on the current short-term data collection plan submitted by the State of North Carolina, the statistical power is computed to be 80.258%. This level of power is considered high and should be adequate to meet the published Fort Bragg Evaluation statement of work."

Two nationally prominent consultants have reviewed Vanderbilt's analysis and the HCSCI report; their responses and copies of their vitae are included in Appendix A. Dr. Jim Mintz at UCLA is an expert in the conduct and analysis of clinical trials, and is a former student of the renowned expert in power analysis, Dr. Jacob Cohen. Dr. Mark Lipsey, a colleague at Vanderbilt, is an expert in power analysis. His most recent book, *Design Sensitivity: Statistical Power for Experimental Research*, is cited by HCSCI's expert consultant, Dr. Kapadia, in her review (which is provided in Appendix C). A technical response to Dr. Kapadia's review is provided in Appendix D.

SUMMARY RESPONSE

This two-page summary response to HCSCI is straightforward and concise; a detailed technical response appears in Appendix B.

1. How did the HCSCI report recommend a smaller sample size than Vanderbilt while keeping power ($\sim .80$) the same?

It used a statistical analysis procedure different from the one proposed by Vanderbilt. The HCSCI *t*-test requires fewer clients to obtain statistical significance

2. Is the analysis correct in the HCSCI report?

No. It proposed a one-tailed posttest only *t*-test. This test assumes two things: (1) that the groups being compared are equivalent before the intervention and (2) it is not possible that the Health Services Command's traditional services could be better than the Rumbaugh services. Because the study is a quasi-experiment, where the clients are not randomly assigned to treatment and control groups, and HCSCI was informed that the treatment and control groups differ significantly at baseline on some variables, a posttest only design is not tenable. It is for this very reason that Vanderbilt designed a study, approved by HCSCI four years ago, that included baseline and Wave 2, as well as Wave 3, data collection. In the plan suggested by the HCSCI report, we would ignore the baseline and Wave 2 data that were collected at a cost of several million dollars.

Availability Codes	
Dist	Availability for Speciation
A-1	

3. *Did the HCSCI report describe what was wrong with the Vanderbilt statistical plan?*

No, the report proposed a different statistic without criticizing the repeated measures analysis proposed by Vanderbilt.

4. *What did the HCSCI report criticize about Vanderbilt's power analysis?*

As noted earlier, it found no fault with the repeated measures analysis chosen by Vanderbilt, but criticized the use of Monte Carlo simulation to judge the power of the repeated measures design. The HCSCI report stated that it was "*an inappropriate application of this type of simulation*" (p. 7). In fact, the report questions the use of any simulation technique since it does not involve "actual data" (p. 8). The HCSCI report evidently does not recognize that their procedure of looking up power in a table is also a simulation that does not involve actual data. The details of this point are discussed in Appendix B.

Since the Evaluation employs a complex design and equally complex statistical analysis, there are no simple "look up" tables to use, as the HCSCI group did, to judge the power of their *t*-test analysis. Instead, Vanderbilt ran the actual analysis on simulated data to obtain the power estimate. The HCSCI report confuses the method used to obtain the power estimate with the results. As noted in Appendix B, when we apply this simulation method to HCSCI's *t*-test, we obtain a similar result. Therefore, the Monte Carlo simulation technique applies to the current problem in a valid way.

5. *What can Vanderbilt conclude from the HCSCI report?*

A. Since the HCSCI report did not criticize Vanderbilt's statistical methods (repeated measures analysis of variance), there must be no problem with that approach. This is an important conclusion, since the power estimate should be based on the analytic method that will be used, not some other procedure.

B. Assuming their technical competence, HCSCI selected an inappropriate statistical method (a one-tailed posttest only *t*-test) that required fewer clients because they wanted to demonstrate to the Department of Defense that fewer clients were needed to conduct the Evaluation.

C. We have wasted a great deal of the government's money and lost a significant amount of time on the power analysis controversy.

INDEX OF APPENDICES

Appendix A.

Letters from Drs. Lipsey and Mintz, Consultants to Vanderbilt Center for Mental Health Policy; Vitae of Drs. Lipsey and Mintz

Appendix B.

Vanderbilt's Technical Response to HCSCI Report CR93-002

Appendix C.

Letter from Dr. Kapadia, Consultant to Health Services Command

Appendix D.

Vanderbilt's Technical Response to Dr. Kapadia's Letter

Appendix E.

HCSCI Report CR93-002

Appendix F.

Health Services Command's May 21, 1993 Request for Data for Independent Power Analysis

Appendix G.

Vanderbilt's Power Analysis of the Evaluation

Appendix A.

Letters from Drs. Lipsey and Mintz,
Consultants to Vanderbilt Center for Mental Health Policy;
Vitae of Drs. Lipsey and Mintz

Review of "Assessment of Two Data Collection Approaches for Fort Bragg Child/Adolescent Mental Health Demonstration Project Using Power Analysis: A Report to the Assistant Secretary of Defense (Health Affairs)" by Barbara E. Wojcik, Catherine R. Stein, & Scott A. Optenberg, June 1993.

By Mark W. Lipsey, Ph.D.

Context

As I understand it, the Fort Bragg Demonstration Project collects a number of measures on two nonrandomly assigned groups ("demonstration" and "comparison") at three points in time (admission, 6 months, 12 months) for the purpose of comparing the groups for differences that may be attributed to the treatment received by the demonstration group.

At issue is whether sufficient statistical power is achieved for this comparison under a "short-term plan" for which 299 demonstration respondents and 150 comparison respondents are expected to complete all three waves of data collection. The Vanderbilt researchers have provided an analysis that indicates that the short-term plan will not yield sufficient power and have proposed a "long-term plan" that is expected to result in data from 426 demonstration respondents and 361 comparison respondents and which is expected to yield adequate power.

The Army report (a) presents an alternative statistical power analysis that indicates that the short-term plan does yield sufficient power and (b) critiques the procedures the Vanderbilt researchers used to conduct their power analysis. This review will examine both these points.

The Army's Power Analysis

Statistical power for testing a difference between two groups on measured values is a function of four elements; given information on these elements, power is, in principle, determined:

1. The effect size to be detected.
2. The number of respondents in each group.
3. The Type I error rate (alpha) stipulated.
4. The statistical test used.

In addition, to judge the adequacy of the resulting power level, one must stipulate the acceptable Type II error (beta) for the statistical inference at issue.

I will examine the differences between the Army and Vanderbilt power analyses by examining each of these elements.

Power threshold. Both the Vanderbilt and Army analyses set power = .80 (beta = .20) as the minimum level judged sufficient for this study. I will comment on this later, but for now only note that the assumptions about this element do not differ between the Vanderbilt and Army analysis.

Effect size. Both the Vanderbilt and Army analyses assume an effect size of .25 standard deviations difference between the demonstration group and comparison group means in the

wave 3 data. This figure was proposed by the Vanderbilt researchers as a reasonable expectation based on a meta-analysis by Lampman, Durlak, and Wells. Assumptions about this element thus do not differ between the Vanderbilt and Army analysis, but I will also want to comment on this later.

Number of respondents. This element also does not differ between the Army and Vanderbilt. Both assume sample sizes of 299 and 150 in the short-term plan and 426 and 361 in the long-term plan.

Type I error (alpha). There is an important difference between the analyses on the alpha level stipulated. All indications I find in the material available to me shows the Vanderbilt analysis based on the conventional two-tailed $\alpha = .05$ as the stipulated statistical significance threshold. The Army analysis uses a one-tailed $\alpha = .05$ level. No explanation is provided for this departure from the conventional level. The implication of this one-directional test, in comparison to the two-directional one, is that the Army is not interested in any test of whether the outcomes for the demonstration group are worse than those for the comparison group, i.e., that the treatment causes harm rather than benefit. I see no justification for this presumption and find it implausible that the possibility of negative treatment effects is not worthy of consideration in this study.

It is worth noting that had the Army analysis used the conventional two-tailed $\alpha = .05$ level their procedure would have yielded an estimated power of about .68 for the short-term plan rather than the .78/.80 value they derive. This is clearly below the .80 level agreed to be sufficient and supports the Vanderbilt claim that the short-term plan does not yield adequate power.

Statistical test. Power analysis is specific to the statistical test to be used. Moreover, different statistical tests may yield different power, even when applied to the same data. The most striking discrepancy between the power analyses of the Army and Vanderbilt is in the statistical test assumed. The Vanderbilt power analysis assumes a repeated measures ANOVA that will incorporate all three waves of data and test treatment effects in the groups x trials interaction term. The Army analysis assumes a simple t-test comparison between only wave 3 data for the two groups. The Army report provides no explanation or justification for basing its power analysis on a different statistical test than the Vanderbilt researchers have planned nor for adopting a procedure that ignores the wave 1 and wave 2 data.

The Army's assumption of a t-test might be justified if power analysis results for that test could be expected to approximate those for the repeated measures ANOVA. That is clearly not the case, however. A t-test on one wave of data is quite different from a repeated measures ANOVA on all three waves. The data is different, the error terms are different, and there is no doubt that the statistical power will be different.

In this instance, the Vanderbilt analysis estimates a power of about .50 for the ANOVA test; the comparable value for the Army analysis is .68 (two-tailed). It is somewhat surprising that the ANOVA yields lower power since in general the repeated measures format provides some variance control (uses a smaller error denominator in the statistical test) relative to the t-test. This advantage, however, appears to be offset in this case by two factors that work to decrease power. First, the treatment effect in the ANOVA must be tested as an interaction term, while it is a main effect in the t-test. Tests of interactions with given n, alpha, etc.

characteristically have less power than tests of main effects (Cohen, 1988, chap. 8). Second, the ANOVA test builds in a comparison of the two groups for all three waves of measures. When the null hypothesis is false (the assumption of a power analysis), the differences between the groups will be smaller at wave 1 than at wave 2, and smaller at wave 2 than wave 3. The t-test on wave 3 examines only the largest difference between groups; the repeated measures ANOVA integrates information on all three waves and hence implies a smaller net main effect and correspondingly lower power.

This raises the question of whether the Army's assumption of a t-test is justified on the grounds that it is a more appropriate (and more powerful) form of statistical analysis for the data at issue than the repeated measures ANOVA proposed by Vanderbilt. I see no merit to that view, however. A t-test on wave 3 data alone degrades the research design and increases the problems of internal validity. The demonstration and comparison groups are not randomly assigned and, additionally, are subject to increasing attrition problems with successive waves. Such circumstances render ambiguous the extent to which wave 3 differences are attributable to treatment. The analysis planned by the Vanderbilt researchers tests the differences between the two groups not only in terms of the gap evident in wave 3 data, but also in terms of the changes from wave to wave for each group. Thus the groups are being compared with their prior status as well as with each other. This is the effect of testing treatment effects as the group x trials interaction in a repeated measures ANOVA. As such, it makes a more probing examination of potential treatment effects and, by looking at the whole pattern of change rather than just the final cross-section, reduces somewhat the problems of internal validity inherent in this type of field experimentation. In short, the repeated measures ANOVA does appear to be the appropriate statistical test for this situation and not the t-test. The slight increase in power associated with the t-test is gained at the expense of a loss of internal validity and interpretability of the results of the study.

Since the Army's statistical power analysis is based on a statistical test that is not appropriate to the study under consideration, the power results it yields are largely irrelevant. What would be relevant would be the Army's independent calculation of the power expected in a test of the interaction term in a three-wave repeated measures ANOVA comparing groups of $n=299$ and 150 . No such calculation is presented in this report.

Summary. The discrepancy between the Army's power analysis of the short-term plan and the Vanderbilt analysis results from two factors: 1) The Army analysis assumes a one-tailed significance test without presenting any justification for that unconventional choice; the Vanderbilt analysis uses the conventional two-tailed procedure. 2) The Army analysis is based on a statistical test that is different from that proposed by Vanderbilt and inappropriate for the nature of the data and the research design at issue; the Vanderbilt analysis assumes an appropriate statistical test procedure. In my opinion, the Army analysis does not properly address the question of the statistical power of the short-term plan and thus its results are not relevant to assessing that plan. Even if taken at face value, the Army's procedure shows that the short-term plan has insufficient statistical power under conventional two-tailed $\alpha=.05$ significance testing.

Power threshold and effect size revisited. Both the Army and Vanderbilt accept the "Cohen convention" of targeting power at a minimum of .80. I would point out that this leaves a 20% probability of obtaining statistically null results when in fact there is a difference between the groups. Where important treatment effects are at issue I see no justification for

holding Type I error to 5% and allowing Type II error to range as high as 20%. As I have argued elsewhere (Lipsey, 1990), where both errors are equally serious there is a logic for trying to hold them to the same level. In this case that would mean either relaxing alpha, which is quite unconventional, or tightening beta, i.e., attempting to ensure power of .95. Against this higher standard, the short-term plan is even more inadequate.

Similarly, I question the stipulation of the effect size at .25, accepted by both the Army and Vanderbilt and based on the Lampman et al. meta-analysis. A better approach is to determine the minimal effect believed practically important in the treatment context at issue. In many situations this is less than .25. An effect size of .20 standard deviations, for example, represents approximately a 10% improvement in the success rate of a treatment group compared to a control. As a rule of thumb, I find that a reasonable effect size to target in cases where there is no stronger framework for setting the minimum. For detecting .20 instead of .25, both the Army and the Vanderbilt analysis would have found lower power in the short-term plan.

Other Power Considerations

In the concluding section of this report, the Army gives brief discussion to some techniques for enhancing power other than increasing sample size, e.g., reducing variance, using continuous rather than dichotomous measures, and improving reliability of measurement. The latter two points, even if applicable to the ongoing data collection for the Fort Bragg project, would likely produce marginal gains at best. It is true, however, that judicious use of any covariates related to individual differences on outcome measures could enhance the statistical power of the short-term plan by reducing the effective role of individual differences variance in the statistical test. The extent of this enhancement would depend upon the magnitude of the correlation of those covariates with the component of respondent outcome that is not correlated with prior wave measures of that same outcome. Since the prior wave "pretest" measures already built into the repeated measures design are likely to be the most powerful covariates, it is doubtful that others are available that are capable of enhancing the statistical power by an appreciable amount, e.g., enough to raise it from .50 to .80 in the short-term plan.

Another consideration is that the power analyses presented by both Vanderbilt and the Army deal only with the aggregate comparison of all demonstration respondents with all comparison respondents. It is likely to also be appropriate for this study to examine treatment effects for selected subgroups of respondents. Such analysis will necessarily involve smaller sample sizes and hence lower statistical power. In such circumstances it is advisable to design the study so that important subgroup comparisons have adequate power in their own right. The consequence of this is that the power for the aggregate comparison will appropriately be greater than the minimum needed to test only at that level. On these grounds also, then, the .80 target power level for the aggregate comparison is arguably too low.

The Army's Critique of the Vanderbilt Power Analysis

As noted above, the Army's power analysis was based on the assumption of a t-test as the basis for statistical analysis. The Vanderbilt power analysis was based on the assumption of a more appropriate statistical procedure, the repeated measures ANOVA. The question here

is how one should go about doing a power analysis for the repeated measures ANOVA, specifically, for the groups x trials interaction term in that ANOVA. The Army critique of the Vanderbilt approach was quite disparaging.

For context it should be noted that it is not a trivial matter to come up with a power analysis for a specific instance of a repeated measures ANOVA analysis more complex than one-group with two-trials. The major reference book on power analysis (Cohen, 1977, 1988) does not address this analysis and provides no tables in which power values can be looked up, nor do other recognized references (Kraemer & Thiemann, 1987; Lipsey, 1990). The technical literature (e.g., Lui & Cumberland, 1992) shows that power for main effects in such designs are complex functions of the effect size, the number of subjects, the number of waves, the ratio of subject individual differences variance to error variance, the reliability of the measures, and the variation of the test instrument itself. For most realistic situations, power for the main effect in this design would have to be estimated with calculus computations directly or approximations using specialized computer algorithms. For interaction effects in such designs, which is where the power issue lies for the Fort Bragg study, there appears to be even less guidance in the literature for practical approaches to power analysis.

The Vanderbilt researchers approached this difficult situation by using a Monte Carlo simulation to produce a large number of artificial data sets that varied by sampling error around the parameters expected to be reflected in the Fort Bragg data-- means, variances, effect size, correlation between waves, etc., based, where possible, on the results found in the actual data from waves 1 and 2. They then tested the statistical significance of the appropriate interaction in each such data set using the repeated measures ANOVA. Since statistical power is defined as the probability of obtaining statistical significance given that the null hypothesis is false, the proportion of times that significance was found in this procedure was used as an estimate of statistical power.

Such Monte Carlo approaches are not uncommon and, appropriately done, are quite credible in statistical work, e.g., are published in reputable statistical journals. I have looked over the printouts and reports that describe the simulation conducted by the Vanderbilt researchers. While I cannot attest that every detail is correct, I find no obvious errors and find their procedure quite reasonable.

The Army report, however, attacks this procedure not by demonstrating any specific error, nor by presenting an alternate analysis derived directly from statistical theory. Rather, it presents selected quotes from approximately 20 year old textbooks in operations research and econometric modeling to the effect that poor simulations yield poor results. The few substantive points are not well developed and are arguable. The major theme is that the parameters of the simulation should be taken from actual data. No mention is made of the fact that some of the parameters of the Vanderbilt simulation were taken from actual data or that sufficient data is not yet in hand to properly estimate all the relevant parameters. Indeed, if such data were in hand power analysis would not be needed. Data with sufficient statistical power to properly estimate the parameters of the simulation would also have sufficient power to estimate the treatment effect directly.

The larger point here is that the task at hand is to do power analysis for planning purposes for a design involving the test of an interaction effect in a three- wave repeated measures

ANOVA. While the Vanderbilt approach may not be optimal, it is reasonable. A proper critique would be to compare the results of that approach with an independent derivation or to point out specific errors that clearly bias the Vanderbilt results. The Army critique does neither.

Conclusion. The power analysis presented by the Army is seriously flawed for purposes of judging the statistical power of the short-term plan. It is biased by an inexplicable assumption of a one-tailed significance test and is based on a statistical testing procedure (t-test) that is not appropriate to the research design of the study at issue. The Vanderbilt power analysis has the advantages of assuming conventional two-tailed testing and attempting to estimate power for the repeated measures ANOVA that is appropriate to the data. While a power analysis based on purely statistical formulations for the pertinent interaction term in that ANOVA would have been preferable, it is not apparent that practical procedures for such an analysis are available in the technical literature. The Monte Carlo simulation that was used instead represents a straightforward and reasonable approach to the problem and its results appear credible. The critique of that simulation presented in the Army report is not substantive and does not provide arguments sufficient to discredit its results.

References

- Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.
- Kraemer, H.C., & Thiernann, S. (1987). How many subjects? Statistical power and analysis in research. Newbury Park, CA: Sage.
- Lipsey, M. (1990). Design sensitivity: Statistical power for experimental research. Newbury Park, CA: Sage.
- Lui, K-J, & Cumberland, W.G. (1992). Sample size requirements for repeated measures in continuous data. Statistics in Medicine, 11, 633-641.

Jim Mintz, Ph.D.
1860 Calle Alberca
Camarillo, CA 93010
(805) 482-3075

July 29, 1993

Warren Lambert, Ph.D.
Center for Mental Health Policy
Vanderbilt University
Nashville, TN 37235

Re: Comments on *Assessment of two data collection approaches for Fort Bragg...* (Wojcik, Stein and Optenberg, 1993)

Dear Dr. Lambert,

I have reviewed the power analyses and critique cited above (referred to below using the acronym *WSO* after the authors' initials) and offer the following reactions:

1. Two questionable assumptions underlie the *WSO* power analyses:

a. The first questionable assumption is that it is appropriate to use the one-tailed .05 alpha criterion. Jacob Cohen, my mentor and dissertation sponsor at NYU in 1969 taught me that clinical research is necessary because our untested beliefs are all too often wrong. Jack taught that the one-tailed test is appropriate only when the researcher is prepared to doggedly attribute results in the "wrong" direction to factors such as error in data collection rather than true experimental effects, regardless of their magnitude. In my opinion, all power analyses in clinical research should be based on the two-tailed .05 alpha level. Substitution of the two-tailed levels results in substantially larger estimated *N*s.

b. The second questionable assumption in the *WSO* analyses is that only one wave of follow-up data should be analyzed. The simple t-test calculations presented in *WSO* are correct as calculated. However, if two waves of data were collected, these uncorrected t-test probability levels would not be correct. One way scientists "adjust" the statistical tests in this situation is the so-called "Bonferroni adjustment" (i.e., use of $p=.025$ rather than .05, because two waves are analyzed rather than one). Another is use of some form of repeated measures or MANOVA analysis, which would be very similar with regard to power estimates. Whichever method one used, more subjects would be needed for adequate power than one would project based on simple one-wave t-tests.

2. Is analysis of both waves of follow-up data desirable?

In my opinion, analysis of both waves of follow-up would be highly desirable. It would

inform us regarding rapidity of improvement, durability of recovery, and a number of issues related to attrition. The price paid by the clinical researcher to look at these questions is a larger sample size. To analyze both waves of follow-up data, we must shift to some kind of repeated measures strategy. The *WSO* analyses are based on t-tests, and they do not deal with the problems of repeated measures analysis. One potentially serious problem, expected in this study, is failure of the assumption of compound symmetry, or equal covariances across time. The Monte Carlo simulations were undertaken to evaluate the seriousness of that problem.

3. In some cases, Monte Carlo approaches may be necessary

I do not agree with *WSO* that Monte Carlo simulations are only valid when they are based on sample data. My view is quite to the contrary, although I certainly agree that simulations on computers must never substitute for thinking. Monte Carlo simulations can, however, tell us about the performance of statistical tests in situations in which we simply do not trust that the Tables are accurate. In the current research situation, the presumption of AR1 structure in the data, in combination with the unequal *N*s, made it impossible to simply look up the power statistics in a Table. In my opinion, there is no reason to doubt the results of the Monte Carlo studies presented. I have previously expressed some of my own reservations regarding the assumptions underlying your computer simulation studies, but my concerns related to the analytic models chosen for study rather than the methodology itself.

In summary, the *WSO* analyses appear to be technically correct. However, they explore power for an entirely different analytic model than the one used in the Vanderbilt Monte Carlo simulation study. The *WSO* analyses assume one wave of follow-up, and use a one-tailed criterion. The simulation studies assume two correlated waves of follow-up, AR1 structure, an interest in analyzing both, and unequal *N*s and covariances. It is not surprising that the latter approach suggests a need for more subjects.

I hope these relatively brief comments are useful. I will be happy to provide you with more extensive feedback on these matters when time permits.

With warm regards, I remain



Jim Mintz, Ph.D.
Professor and Chief,
Methodology and Statistical Support Unit
MHCRC for the Study of Schizophrenia, UCLA
Los Angeles, CA

Curriculum Vitae

MARK W. LIPSEY

Personal

Office address: Department of Human Resources, Box 90 GPC
Vanderbilt University, Nashville, TN 37203

Home address: 303 Fairfax Ave., Nashville, TN 37212

Telephone: Office (615) 343-1586; Home (615) 292-0992

Date of birth: May 8, 1946

Family status: Married; two children

Social Security #: 412-76-3865

Education

Ph.D. (Psychology), The Johns Hopkins University, 1972

B.S. (Applied Psychology), Georgia Institute of Technology, 1968

Employment History

1992- Professor of Public Policy, Department of Human Resources,
Vanderbilt University

1984-92 Professor, Psychology Department, Claremont Graduate School

1984-89 Chairman, Psychology Department, Claremont Graduate School

1978-84 Associate Professor, Claremont Graduate School

1972-78 Assistant Professor, Claremont Graduate School

1971-72 Part-time Instructor, University of Maryland, Baltimore County Campus

Areas of Specialization and Current Interest

Social intervention; program evaluation research:

Treatment effectiveness research
Human service organizations
Juvenile delinquency programs
Quality of life

Applied research methodology:

Experimental/quasi-experimental design and analysis
Meta-analysis
Applied measurement
Survey research

Professional Memberships

American Psychological Association (Divisions 8, 9, 27)
American Psychological Society
International Association of Applied Psychology
American Evaluation Association
American Society of Criminology
Sigma Xi
American Association of University Professors

Professional Activities and Awards

Editorial Board, *Evaluation and Program Planning*, 1992-1995.
Visiting Fellow, U.S. General Accounting Office, Program Evaluation and Methodology Division, 1991-92.
Board Member, Association for Criminal Justice Research, California, 1990-92.
Research Advisory Group, California Department of Mental Health, Sex Offender Treatment and Evaluation Project, 1989-.
Advisory Committee, Foster Family-Based Treatment Association, Minneapolis, 1989-.
Editorial Advisory Board, *New Directions for Program Evaluation*, 1989-1995
Pathways Panel, Program on Human Development and Criminal Behavior, McArthur Foundation and the National Institute of Justice, 1988.
Editorial Advisory Board, *Evaluation Studies Review Annual*, 1987.
Associate Editor, *Evaluation Review*, 1985-1989
Fulbright Lecturer, University of Delhi, India, 1985-86.
Editor-in-Chief, *New Directions for Program Evaluation*, the journal of the American Evaluation Association and a Jossey-Bass Sourcebook, 1984-88
Elected member, Communications Committee, Publication and Communications Board, American Psychological Association, 1973-76 (Chairman, 1976).
NSF Graduate Traineeship in Psychology, The Johns Hopkins University, 1968-1972.

Consulting Experience (Organizations/Projects)

Los Angeles County Probation Dept.	Health Maintenance Network of
Claremont League of Women Voters	Southern California (Blue Cross)
Coldwell Banker Commercial Group	Claremont Civic Association
Tri-Cities Mental Health Authority	American Psychological Association
Shye Research Enterprises, Inc.	Arthur D. Little, Inc.
Los Angeles County Department of	Cerritos Corridor Diversion Project
Community Development	Positive Alternatives for Youth
Los Angeles County Youth Services	Diversion Project (PAY)
Network	
Lewin and Associates, Inc.	Southeast Early Diversion Project (SEED)
West San Gabriel Valley Juvenile	Orange County Criminal Justice
Diversion Project	Planning Council
Pomona Valley Youth Services	Riverside County Department of Public
Project	Social Services
HEAVY-West Youth Services Project	Control Data Corporation

(Continued)

Consulting Experience (Organizations/Projects) (Continued)

HEAVY-San Fernando Valley Youth
Services Project
Los Angeles County Justice System
Advisory Group (JSAG)
Notre Dame Project on Values and
the Electric Power Industry
Claremont Unified School District
National Institute of Alcoholism
and Alcohol Abuse
Kaiser Permanente Medical Care
Program
National Institute of Mental Health

Los Angeles County Superintendent
of Schools Office
Los Angeles County Regional Criminal
Justice Planning Board
Orange County Grand Jury
Navy Personnel Research and
Development Center
Risk/Need Project, Ministry of Community
and Social Service, Ontario, Canada
U.S. General Accounting Office Program Evaluation
and Methodology Division
Tennessee Commission on Children and Youth

Research Grants and Contracts

Russell Sage Foundation: "A Meta-analysis of Juvenile Delinquency Treatment Effectiveness Research,"
1987-89 (\$44,261).
National Institute of Mental Health: "Meta-analysis of Juvenile Delinquency Treatment Research; MH42694,
1987-89 (\$82,000).
Orange County Grand Jury: "Organizational Study of the Government of the County of Orange," 1987
(\$30,200).
National Institute of Mental Health: "Meta-analysis of Juvenile Delinquency Treatment Evaluation Research,"
MH39958, 1985-87 (\$114,018).
National Institute of Justice: "Measurement Issues in the Evaluation of the Effects of Juvenile Justice
Programs," 80-IJ-CX-0036, 1980-81 (\$28,704).
Los Angeles County: "1981 Juvenile Justice Program Evaluation (Youth Services Network)," LARCJPB
Agreement #81-82, 1981-82 (\$100,000).
Los Angeles Regional Criminal Justice Planning Board: "1980 Juvenile Justice Program Evaluation (Youth
Services Network)," LARCJPB Agreement #80-1, 1980-81 (\$59,897).
Los Angeles County, County Justice System Subvention Program: "Countywide Impact of Juvenile Diversion
Network," Project No. 34150, 1979 (\$17,666).
Miscellaneous contracts for program evaluation with individual youth service projects, 1977-81 (\$25,690 total).
Orange County, California: "Regional Diversion Program Evaluation, 1977-78 (\$114,771).
Los Angeles County Sheriff's Department Juvenile Diversion Programs: Evaluation research, 1976-79 (\$94,000
total).
NIMH Small Grant, "Follow-up Study of Psychologists." MH23474, 1974-75 (\$5,692).
Control Data Corporation, Los Angeles, California: Survey of social services in Riverside County, 1973
(\$25,000).
NSF Grant for Dissertation Research: "Scientific Knowledge and Scientific Norms in Psychology. #GS-30891,
1971-72 (\$5,000).
APA: "Survey of Graduate Students in Psychology," #70-5, 1970-71 (\$10,000).

Books

Mark W. Lipsey. *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage Publications, 1990.

David S. Cordray & Mark W. Lipsey (Eds.). *Evaluation Studies Review Annual, Volume 11*. Beverly Hills, CA: Sage Publications, 1987.

Articles and Chapters

Mark W. Lipsey. Identifying potential variables and analysis opportunities. Chapter 9 in H. Cooper & L.V. Hedges (Eds.). *The handbook of research synthesis*. NY: Russell Sage Foundation (in press for 1993).

Mark W. Lipsey & Robert A. Keith. The importance of theory in assessing the quality of rehabilitation services. In R. Glueckauf, L. Sechrest, G. Bond, & J. McGrew (Eds.). *Improving the quality of assessment practices in rehabilitation psychology*. NY: Pergamon, 1992 (in press).

Mark W. Lipsey. *The effect of treatment on juvenile delinquents: Results from meta-analysis*. In F. Losel, D. Bender, & T. Bliesener (Eds.). *Psychology and law: International perspectives* (pp. 131-143). Berlin; NY: Walter de Gruyter.

Mark W. Lipsey. Meta-analysis in evaluation research: Moving from description to explanation. In H.T. Chen & P.H. Rossi (Eds.). *Using theory to improve program and policy evaluations* (pp. 229-241). NY: Greenwood Press, 1992.

Mark W. Lipsey. Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In T.D. Cook, H. Cooper, D.S. Cordray, H. Hartmann, L.V. Hedges, R.J. Light, T.A. Louis, & F. Mosteller (Eds.). *Meta-analysis for explanation: A casebook*. NY: Russell Sage Foundation, 1992.

Joseph A. Durlak & Mark W. Lipsey. A practitioner's guide to meta-analysis. *American Journal of Community Psychology*, 1991, 19(3), 291-332.

Mark W. Lipsey. Book review of: H.T. Chen, *Theory-driven evaluations*. *Evaluation and Program Planning*, 1991, 14, 412-414.

Mark W. Lipsey. Book review of: A.R. Roberts, *Juvenile justice: Policies, programs, and services*. *Contemporary Psychology*, 1990, 35(12), 1174.

Mark W. Lipsey. Theory as method: Small theories of treatments. In L. Sechrest, E. Perrin, & J. Bunker (Eds.). *Research methodology: Strengthening causal interpretations of nonexperimental data*. Washington, D.C.: U.S. Public Health Service, Agency for Health Care Policy and Research, 1990.

Mark W. Lipsey. Core curriculum: An idea whose time has passed. In L. Bickman & H. Ellis (Eds.). *Preparing psychologists for the 21st century: Proceedings of the National Conference on Graduate Education in Psychology*. Hillsdale, NJ: Lawrence Erlbaum, 1990.

Mark W. Lipsey & John A. Pollard. Driving toward theory in program evaluation: More models to choose from. *Evaluation and Program Planning*, 1989, 12, 317-328.

Mark W. Lipsey. Practice and malpractice in evaluation research. *Evaluation Practice*, 1988, 9(4), 5-24.

- Mark W. Lipsey. Juvenile delinquency intervention. *New Directions for Program Evaluation*, 1988, 37, 63-84.
- David S. Cordray & Mark W. Lipsey. Evaluation studies for 1986: Program evaluation and program research. *Evaluation Studies Review Annual*, 1987, 11, 17-44.
- Mark W. Lipsey, Scott Crosse, Jan Dunkle, John Pollard, & Gordon Stobart. Evaluation: The state of the art and the sorry state of the science. In D.S. Cordray (Ed.). Utilizing prior research in evaluation planning. *New Directions for Program Evaluation*, 1985, 27, 7-28.
- Mark W. Lipsey. Book review of: L. Saxe & M. Fine, *Social experiments: Methods for design and evaluation*. *Evaluation and Program Planning*, 1984, 7, 387-388.
- Georgine M. Pion & Mark W. Lipsey. Psychology and society: The challenge of change. *American Psychologist*, 1984, 39(7), 739-754.
- Mark W. Lipsey. Is delinquency prevention a cost-effective strategy? A California perspective. *Journal of Research in Crime and Delinquency*, 1984, 21, 279-302.
- Mark W. Lipsey. Program evaluation research. In R.J. Corsini (Ed.). *Wiley encyclopedia of psychology*. NY: John Wiley, 1984.
- Mark W. Lipsey. A scheme for assessing measurement sensitivity in program evaluation and other applied research. *Psychological Bulletin*, 1983, 94, 152-165. (Reprinted in R. Conner, et al. (Eds.). *Evaluation Studies Review Annual*, Volume 9. Beverly Hills, CA: Sage, 1984.)
- Teris K. Schery & Mark W. Lipsey. Program evaluation for speech and hearing services. In J. Miller, R. Schiefelbusch, & D. Yoder (Eds.). *Language intervention*. Trenton, NJ: B.C. Decker, Inc., 1982.
- Robert B. Huebner & Mark W. Lipsey. The relationship of three measures of locus of control to environmental activism. *Basic and Applied Social Psychology*, 1981, 2(1), 45-58.
- Mark W. Lipsey, David S. Cordray, & Dale E. Berger. Evaluation of a juvenile diversion program: Using multiple lines of evidence. *Evaluation Review*, 1981, 5(3), 283-306.
- Georgine M. Pion & Mark W. Lipsey. Public attitudes toward science and technology: What have the surveys told us? *Public Opinion Quarterly*, 1981, 45, 303-316. (Reprinted in J.E. Twining (Ed.). *Reading and thinking: A process approach*. NY: Holt, Rinehart, & Winston, 1984).
- Raymond Coffin & Mark W. Lipsey. Moving back-to-the-land as an ecologically responsible lifestyle change. *Environment and Behavior*, 1981, 13(1), 42-63.
- Mark W. Lipsey. Occupational socialization and mid-career orthodoxy among academic psychologists. *Personality and Social Psychology Bulletin*, 1978, 4(1), 169-172.
- Richard W. Anderson & Mark W. Lipsey. Energy conservation and attitudes toward technology. *Public Opinion Quarterly*, 1978, 42(1), 17-30.
- Mark W. Lipsey. Adaptation and the technological society: A value context for technology assessment. *Zygon*, 1978, 13(1), 2-18.

Dan Perlman & Mark W. Lipsey. Who's who in social psychology: A textbook definition. *Personality and Social Psychology Bulletin*, 1978, 4(2), 212-216.

Mark W. Lipsey. Attitudes toward the environment and pollution. In S. Oskamp. *Attitudes and opinions*. Englewood Cliffs, NJ: Prentice-Hall, 1977, pp. 360-379.

Mark W. Lipsey. The personal antecedents and consequences of ecologically responsible behavior: A review. *JSAS Catalog of Selected Documents in Psychology*, 1977, 7, 70 (36 pages).

Arthur H. Brayfield & Mark W. Lipsey. Public affairs psychology. In P.J. Woods (Ed.). *Career opportunities for psychologists*. Washington, D.C.: American Psychological Association, 1976.

Mark W. Lipsey & Arthur H. Brayfield. *Profession of psychology. Programmed learning aid, Introductory psychology series*. Homewood, Illinois: Richard D. Irwin, Inc., 1975.

Mark W. Lipsey. Psychology: Preparadigmatic, postparadigmatic, or misparadigmatic? *Science Studies*, 1974, 4, 406-410.

Mark W. Lipsey. Research and relevance: A survey of graduate students and faculty in psychology. *American Psychologist*, 1974, 29(7), 541-553.

Mark W. Lipsey. Psychology tomorrow: A survey of graduate students and faculty in psychology. *JSAS Catalog of Selected Documents in Psychology*, 1973, 3, 36 (129 pages).

Joseph Sonnefeld & Mark Lipsey. NISP, students, and the open society (Comment). *American Psychologist*, 1972, 27, 340-341.

Mark W. Lipsey. Scientific values and scientific knowledge: A test of an evolutionary model. *JSAS Catalog of Selected Documents in Psychology*, 1972, 2, 113 (97 pages). Published version of doctoral dissertation, *Scientific knowledge and scientific norms in psychology*, 1972.

Invited Addresses (Last Five Years)

Mark W. Lipsey. *What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents?* Plenary presentation, Conference on "What works 1992: Next steps ...," Salford University, Manchester, Great Britain, September 1992.

Mark W. Lipsey. *Social responsibility, mentoring, and research training*. Invited presentation, Conference on Maintaining and Promoting Scientific Integrity in Behavioral Science Research sponsored by the Science Directorate of the American Psychological Association, Vanderbilt Institute for Public Policy Studies, and Vanderbilt University. Vanderbilt University, Nashville, October 1991.

Mark W. Lipsey. *Rehabilitative treatment for delinquents: Meta-analysis, methodology, and myth*. Keynote address, Association for Criminal Justice Research (California). Sacramento, April 1991.

Mark W. Lipsey. *The fuzzy world hypothesis and other lessons from applied research*. Keynote address, California Association for Institutional Research. Sacramento, November 1990.

Mark W. Lipsey. *Meta-analysis and the fuzzy world hypothesis*. Invited presentation, General Government Division, U.S. General Accounting Office. Washington, DC, October 1990.

Mark W. Lipsey & Robert A. Keith. *The importance of theory in assessing the quality of rehabilitation services*. Invited presentation, Rehabilitation Psychology Conference: Improving the Quality of Assessment Practices. Indiana University-Purdue University at Indianapolis, October 1990.

Mark W. Lipsey. *The effects of treatment on juvenile delinquents: Results from meta-analysis*. Plenum address, European Conference on Law and Psychology. Nurnberg, Germany, September 1990.

Mark W. Lipsey. *New developments in program evaluation*. Invited presentation, Human Sciences Research Council. Pretoria, South Africa, April 1990.

Mark W. Lipsey. *Social intervention*. Workshop sponsored by the Department of Industrial Psychology, University of Bophuthatswana, Mmbatho, Bophuthatswana, Africa, April 1990.

Mark W. Lipsey. *Program evaluation*. Workshop sponsored by the Human Sciences Research Council, Pretoria and Capetown, South Africa, April 1990.

Presentations (Last Five Years)

Mark W. Lipsey. *What meta-analysis teaches us about methods for studying delinquency treatment*. The American Society of Criminology. San Francisco, November 1991.

Mark W. Lipsey. *Prediction, prevention, programming, and meta-analysis*. American Evaluation Association. Chicago, October 1991.

Mark W. Lipsey. *Meta-analysis: Patterns of relationships among variables*. American Evaluation Association. Washington, DC, October 1990.

Mark W. Lipsey. *What works in delinquency treatment: Results from 400 studies*. Academy of Criminal Justice Sciences. Denver, March 1990.

Mark W. Lipsey. *The efficacy of intervention for juvenile delinquency: Results from 400 studies*. The American Society of Criminology. Reno, November 1989.

Mark W. Lipsey. *The role of method in determining outcome: Lessons from a meta-analysis of juvenile delinquency treatment*. The American Evaluation Association. San Francisco, October 1989.

Mark W. Lipsey. *Lessons to be learned from 200 treatment effectiveness meta-analyses*. Second Biennial Conference on Community Research and Action. Michigan State University, East Lansing MI, June 1989.

Mark W. Lipsey. *Treatment effectiveness research: Insights from meta-analysis*. American Evaluation Association. New Orleans, October 1988.

Research Reports

Karen Hult, Mark Lipsey, Joseph Maciariello, & Vijay Sathe. *The organizational effectiveness of county government in Orange County*. Research Report, March 1987.

Mark W. Lipsey, John A. Pollard, & Anne Gowan. *Indicators of market strength in commercial real estate: A feasibility study*. Research Report, May 1984.

Mark Lipsey, Allan Wicker, Elyce Kerce, Scott Crosse, Jim Griffith, & Shirley Trosino. *The Health Net constituencies: A survey of the Health Net publics and their perceptions of Health Net performance*. Research Report, July 1983.

Mark W. Lipsey, Jack I. Mills, & Mary Ann Plant. *1980 Los Angeles County Youth Services Network Evaluation*. Research Report, May 1982.

Mark W. Lipsey. *Measurement issues in the evaluation of the effects of juvenile delinquency programs*. Research Report on Project 80-U-CX-0036, National Institute of Justice, Office of Research and Evaluation Methods, June 1982. (National Criminal Justice Reference Service Document No. NCJ-84968).

Mark W. Lipsey, Jack I. Mills, Raymond Coffin, Kathleen B. Fraser, & Mary Ann Plant. *1980 Los Angeles County Youth Services Network Evaluation*. Research Report, July 1981.

Mark W. Lipsey & Judith E. Johnston. *The impact of juvenile diversion in Los Angeles County: A report to the Los Angeles County (AB90) Justice System Advisory Group*. Research Report, July 1979.

Mark W. Lipsey, Dale E. Berger, et al. *Final evaluation report: Southeast Early Diversion Project*. Research Report, December 1978.

Mark W. Lipsey, Dale E. Berger, et al. *Final evaluation report: Cerritos Corridor Diversion Project*. Research Report, December 1978.

Mark W. Lipsey, Dale E. Berger, et al. *Final evaluation report: Positive Alternatives for Youth Diversion Project*. Research Report, December 1978.

Mark W. Lipsey, Dale E. Berger, et al. *Final evaluation report for the Orange County Regional Diversion Program*. Research Report, March 1978.

Mark W. Lipsey, Dale E. Berger, Janet M. Lange, & Laura B. Dennison. *Second annual evaluation report: Southeast Early Diversion Project*. Research Report, January 1978.

Mark W. Lipsey, Dale E. Berger, Janet M. Lange, & Laura B. Dennison. *Second annual evaluation report: Cerritos Corridor Diversion Project*. Research Report, January 1978.

Mark W. Lipsey, Dale E. Berger, Janet M. Lange, & Laura B. Dennison. *Second annual evaluation report: Positive Alternatives for Youth Diversion Project*. Research Report, January 1978.

Mark W. Lipsey, Dale E. Berger, et al. *Interim evaluation report for the Orange County Regional Juvenile Diversion Program*. Research Report, September 1977.

Dale E. Berger, Mark W. Lipsey, Laura B. Dennison, & Janet M. Lange. *The effectiveness of the Sheriff's Department's juvenile diversion projects in southeast Los Angeles County*. Research Report, April 1977.

Mark W. Lipsey, Dale E. Berger, & Laura B. Dennison. *First annual evaluation report: Southeast Early Diversion Project*. Research Report, January 1977.

Mark W. Lipsey, Dale E. Berger, & Laura B. Dennison. *First annual evaluation report: Cerritos Corridor Diversion Project*. Research Report, January 1977.

Mark W. Lipsey, Dale E. Berger, & Laura B. Dennison. *First annual evaluation report: Positive Alternatives for Youth Diversion Project*. Research Report, January 1977.

Mark W. Lipsey & Dale E. Berger. *Riverside County social services survey: Final report*. Research report under contract to Control Data Corporation, October 1974.

SoGSiP Study Group (Mark W. Lipsey, Chairman). *Some preliminary results from a survey of graduate students in psychology*. National Information System for Psychology/American Psychological Association, Technical Report No. 15, March 1971.

CURRICULUM VITA

Jim Mintz, Ph.D.

July, 1993

PRESENT POSITIONS:

Professor
Department of Psychiatry
Division of Medical Psychology
School of Medicine
University of California, Los Angeles

Chief, Methodology and Statistical Services Unit
Clinical Research Center for the Study of Adult
Schizophrenia
UCLA Department of Psychiatry

Research Psychologist
West Los Angeles VA Medical Center
Brentwood Division

BUSINESS ADDRESS:

West Los Angeles VA Medical Center
Brentwood Division (691/B117)
11301 Wilshire Blvd.
Los Angeles, CA 90074
Telephone: 310/477-7927

HOME ADDRESS:

1860 Calle Alberca
Camarillo, California 93010
Telephone: 805/482-3075

EDUCATION:

1962 A.B.
Drew University
Madison, New Jersey

1967 M.S.
School of Education
City College of New York
New York, New York

1969 Ph.D., (Clinical Psychology)
Graduate School of Arts and Sciences
New York University
New York, New York

**PSYCHOLOGY
INTERNSHIP:**

1965-66	VA Hospital East Orange, New Jersey
1966	VA Psychiatric Hospital Palo Alto, California
1966-67	VA Regional Office Mental Hygiene Clinic Newark, New Jersey

**FACULTY
APPOINTMENTS:**

1968-78	Department of Psychiatry School of Medicine University of Pennsylvania Philadelphia, Pennsylvania 1968-69 Assistant Instructor 1969-71 Instructor 1971-72 Associate 1972-77 Assistant Professor 1977-78 Associate Clinical Professor
1978-	Department of Psychiatry, Division of Medical Psychology, School of Medicine, University of California, Los Angeles
1978-81	Associate Professor
1981-	Professor

**PROFESSIONAL
SOCIETIES:**

Society for Psychotherapy Research
President (1985-86)
American Statistical Association

**LICENSURE
(Psychologist):**

California (PL5517)
Pennsylvania (PS002491)

**HOSPITAL
APPOINTMENTS:**

1971-77	Clinical Psychologist, VA Medical Center, Philadelphia, Pennsylvania
---------	---

	1977-81	Clinical Psychologist, West Los Angeles VA Medical Center, Brentwood Division, Los Angeles, California
	1981-	Member, Mental Health Group Practice Plan, Neuropsychiatric Institute, University of California, Los Angeles Los Angeles, California
CONSULTING & OTHER ACTIVITIES:		
	1991-	Research Advisory Panel Aging CRC, Stanford (Yesavage)
	1991-	Research Advisory Panel NIDA Medication Development Unit (Marder)
	1991-	Advisor, Task Force on Prevention (Liberman) NAS
	1992-	Senior Methodological Consultant Major Disorders of Childhood Program Project (Cantwell)
	1991-	Methodological Consultant CA Dept of Mental Health Evaluation Project (Lewin Associates)
	1991-	Professional Staff Committee Neuropsychiatric Institute, UCLA
	1990-	Chair, Quality Assurance Task Force CRC for Schizophrenia, UCLA (Liberman)
	1981-88	Research consultant, West Los Angeles VA Medical Center, Brentwood Division, Los Angeles, California
	1981-88	Consultant to Rehabilitation Medicine Service on program evaluation, West Los Angeles VA Medical Center, Los Angeles, California

- 1981-85 Assessment consultant, NIMH Collaborative
Research Project on Treatments of Depression,
NIMH, Rockville, Maryland
- 1985-88 Member, Treatment Development and
Assessment Review Committee, Program
Project and Clinical Research Centers
Subcommittee (TDAC), NIMH, Rockville,
Maryland
- 1986-87 Consultant to Extramural Policy Branch,
Division of Extramural Review, NIMH,
Rockville, Maryland

ORIGINAL PUBLICATIONS

- Mintz, J. Survey of student therapists' attitudes toward psychodiagnostic reports. *Journal of Consulting and Clinical Psychology*, 32: 500, 1968.
- Mintz, J. A correlational method for the investigation of systematic trends in serial data. *Educational and Psychological Measurement*, 30: 575-578, 1970.
- Mintz, J., Luborsky, L. P-technique factor analysis in psychotherapy research: An illustration of a method. *Psychotherapy: Theory, research and practice*, 7: 13-18, 1970.
- Bachrach, H., Mintz, J., Luborsky, L. On empathy and other psychotherapy variables: An experience with the effects of training. *Journal of Consulting and Clinical Psychology*, 36(3): 445, 1971.
- Mintz, J., Luborsky, L. Segments vs. whole sessions: Which is the better unit for psychotherapy research? *Journal of Abnormal Psychology*, 78(2): 180-191, 1971.
- Mintz, J., Luborsky, L., Auerbach, A. Dimensions of psychotherapy: A factor analytic study of ratings of psychotherapy sessions. *Journal of Consulting and Clinical Psychology*, 36: 106-120, 1971.
- Mintz, J. What is success in psychotherapy? *Journal of Abnormal Psychology*, 80(1): 11-19, 1972.
- Mintz, J., Wiedemann, C. Intraclass correlation as a reliability check on nominal data. *Educational Psychological Measurement*, 32: 801-804, 1972.
- O'Brien, C., Hamm, K., Ray, B., Pierce, J., Luborsky, L., Mintz, J. Group vs. individual psychotherapy with schizophrenics: A controlled outcome study. *Archives of General Psychiatry*, 27: 474-478, 1972.
- Mechanick, P., Mintz, J., Gallagher, J., Lapid, G., Rubin, R., Good, J. Non-medical drug use among medical students. *Archives of General Psychiatry*, 29: 48-50, 1973.
- Mintz, J., Auerbach, A., Luborsky, L., Johnson, M. Patients', therapists' and observers' views of psychotherapy: A 'Rashomon' experience or a reasonable consensus? *British Journal of Medical Psychology*, 46: 83-89, 1973.
- Bachrach, H., Mintz, J. The Wechsler Memory Scale as a tool of detection of mild cerebral dysfunction. *Journal of Clinical Psychology*, 304: 58-60, 1974.

- Luborsky, L., Mintz, J. What sets off momentary forgetting during a psychoanalysis? Investigations of symptom onset conditions. *Psychoanalysis and Contemporary Science*, 3: 233-268, 1974.
- Mintz, J., O'Hare, K., O'Brien, C., Goldschmidt, J. Sexual problems of heroin addicts. *Archives of General Psychiatry*, 31: 700-703, 1974.
- Wiedemann, C., Mintz, J. Student therapists' assessment of diagnostic testing. *Journal of Personality Assessment*, 38(3): 203-214, 1974.
- Luborsky, L., Crabtree, L., Curtis, H., Ruff, G., Mintz, J. The concept space of transference for eight psychoanalysts. *British Journal of Medical Psychology*, 48: 65-70, 1975.
- Mintz, J., O'Hare, K., Goldschmidt, J., O'Brien, C. Double-blind detoxification of methadone maintenance patients. *International Journal of the Addictions*, 10(5): 815-824, 1975.
- O'Brien, C., Greenstein, R., Mintz, J., Woody, G. Clinical experience with naltrexone. *American Journal of Drug and Alcohol Abuse*, 2(3): 365-377, 1975.
- O'Brien, C., O'Brien, T., Mintz, J., Brady, J. Conditioning of narcotic abstinence symptoms in human subjects. *Drug and Alcohol Dependence*, 1: 115-123, 1975.
- Woody, G., Mintz, J., O'Hare, K., O'Brien, C. Diazepam use by patients in a methadone program - how serious a problem? *Journal of Psychedelic Drugs*, Oct-Dec, 7(4): 373-379, 1975.
- Woody, G., O'Hare, K., Mintz, J., O'Brien, C. Rapid intake: A method for increasing retention rate of heroin addicts seeking methadone treatment. *Comprehensive Psychiatry*, Mar-Apr, 16(2): 165-169, 1975.
- Luborsky, L., Mintz, J., Brightman, V., Katcher, A. Herpes simplex virus and moods: A longitudinal study. *Journal of Psychosomatic Research*, 20(6): 543-548, 1976.
- Mintz, J., O'Brien, C., Luborsky, L. Predicting the outcome of psychotherapy for schizophrenics. *Archives of General Psychiatry*, 33: 1183-1186, 1976.
- Stanton, M., Mintz, J., Franklin, R. Drug flashbacks II: Some additional findings. *International Journal of the Addictions*, 11(1): 53-69, 1976.
- Nace, E., Meyers, A., O'Brien, C., Ream, N., Mintz, J. Depression in veterans two years after Vietnam. *American Journal of Psychiatry*, 134: 167-170, 1977.

- Khatami, M., Mintz, J., O'Brien, C. Biofeedback mediated relaxation in narcotic addicts. *Behavior Therapy*, 9(5): 968-969, 1978.
- Mintz, J., O'Brien, C., Pomerantz, B. The impact of Vietnam service on heroin-addicted veterans. *American Journal of Drug and Alcohol Abuse*, 6(1): 39-52, 1979.
- Grabowski, J., O'Brien, C., Mintz, J. Increasing the likelihood that consent is informed. *Journal of Applied Behavior Analysis*, 12(2): 283-284, 1979.
- Luborsky, L., Mintz, J., Christoph, P. Are psychotherapeutic changes predictable? Comparison of a Chicago Counseling Center Project with a Penn Psychotherapy Project. *Journal of Consulting and Clinical Psychology*, 47(3): 469-473, 1979.
- Mintz, J., Luborsky, L., Christoph, P. Measuring the outcomes of psychotherapy: Findings of the Penn Psychotherapy Project. *Journal of Consulting and Clinical Psychology*, 47(2): 319-334, 1979.
- Mintz, J., O'Brien, C., Woody, G., Beck, A. Depression in treated narcotic addicts, ex-addicts, nonaddicts, and suicide attempters: Validation of a very brief depression scale. *American Journal of Drug and Alcohol Abuse*, 6(4): 385-396, 1979.
- Luborsky, L., Mintz, J., Auerbach, A., Christoph, P., Bachrach, H., Todd, T., Johnson, M., Cohen, M., O'Brien, C. Predicting the outcome of psychotherapy: Findings of the Penn Psychotherapy Project. *Archives of General Psychiatry*, 37: 471-481, 1980.
- Mintz, J., Christoph, P., O'Brien, C., Snedeker, M. The impact of the interview method on reported symptoms of narcotic addicts. *International Journal of the Addictions*, 15(4): 597-604, 1980.
- O'Brien, C., Nace, E., Mintz, J., Meyers, A., Ream, N. Follow-up of Vietnam Veterans. I. Relapse to drug use after Vietnam service. *Drug and Alcohol Dependence*, 5(5): 333-340, 1980.
- Nace, E.P., O'Brien, C.P., Mintz, J., Meyers, A.L., Ream, N. Follow-up of Vietnam veterans II. Social adjustment. *Drug and Alcohol Dependence*, 6(4): 209-214, 1980.
- Woody, G., Tennant, F., McLellan, A., O'Brien, C., Mintz, J. Lack of toxicity of high dose propoxyphene napsylate when used for maintenance treatment of addiction. *Clinical Toxicology*, 16(4): 473-478, 1980.

- Mintz, J. Measuring outcome in psychodynamic psychotherapy. *Archives of General Psychiatry*, 38: 503-506, 1981.
- Mintz, J., Steuer, J., Jarvik, L. Psychotherapy with depressed elderly patients: Research considerations. *Journal of Consulting and Clinical Psychology*, 49(4): 542-548, 1981.
- Woody, G.E., Mintz, J., Tennant, F., O'Brien, C.P., McLellan, A.T., Marcovici, M. Propoxyphene for maintenance treatment of narcotic addiction. *Archives of General Psychiatry*, 38(8): 898-900, 1981.
- Woody, G., O'Brien, C., McLellan, A., Mintz, J. Psychotherapy for opiate addiction: Some preliminary results. *Annals New York Academy of Sciences*, 362: 91-100, 1981.
- Jarvik, L.F., Mintz, J., Steuer, J., Gerner, R. Treating geriatric depression: A 26-week interim analysis. *Journal of the American Geriatrics Society*, 30: 713-717, 1982.
- Khatami, M., Woody, G., O'Brien, C., Mintz, J. Biofeedback treatment of narcotic addiction: A double-blind study. *Drug and Alcohol Dependence*, 9(2): 111-117, 1982.
- Jarvik, L.F., Read, S.L., Mintz, J., Neshkes, R.E. Pretreatment orthostatic hypotension in geriatric depression: Predictor of response to imipramine and doxepin. *Journal of Clinical Psychopharmacology*, 3(6): 368-372, 1983.
- Mintz, J. Integrating research evidence: a commentary on meta-analysis. *Journal of Consulting and Clinical Psychology*, 51(1): 71-75, 1983.
- Mintz, J., Jarvik, L.F. Sex differences in brain atrophy during aging (letter) *Journal of the American Geriatric Society*, 31(3): 187-189, 1983.
- Doane, J.A., Falloon, L.R.H., Goldstein, M.J., Mintz, J. Parental affective style and the treatment of schizophrenia: Predicting course of illness and social functioning. *Archives of General Psychiatry*, Jan, 42: 34-42, 1984.
- Marder, S., Van Putten, T., Mintz, J., McKenzie, J., Lebell, M., Faltico, G., May, P.R.A. Costs and benefits of two doses of fluphenazine decanoate. *Archives of General Psychiatry*, Nov, 41: 1025-1029, 1984.
- Steuer, J.L., Mintz, J., Hammen, C.L., Hill, M., Jarvik, L.F., McCarley, T., Motoike, P., Rosen, R. Cognitive-behavioral and psychodynamic group psychotherapy in treatment of geriatric depression. *Journal of Consulting and Clinical Psychology*, 52(2): 180-189, 1984.

- Irwin, M., Lovitz, A., Marder, S.R., Mintz, J., Winslade, W.J., Van Putten, T., Mills, M.J. Psychotic patients' understanding of informed consent. *American Journal of Psychiatry*, Nov, 142(11): 1351-4, 1985.
- Mintz, J., Boyd, G., Rose, J.E., Charuvastra, V.C., Jarvik, M.E. Alcohol ncreases cigarette smoking: a laboratory demonstration. *Addictive Behaviors*, 10(3): 203-7, 1985.
- Mintz, J., Mintz, L.I., Jarvik, L.F. Cognitive behavioral therapy in geriatric depression: reply to Riskind, Beck, and Steer. *Journal of Consulting and Clinical Psychology*, Dec, 53(6): 946-7, 1985.
- Neshkes, R.E., Gerner, R., Jarvik, L.F., Mintz, J., Joseph, J., Linde, S., Aldrich, J., Conolly, M.E., Rosen, R., Hill, M.A. Orthostatic effect of imipramine and doxepin in depressed geriatric outpatients. *Journal of Clinical Psychopharmacology*, Apr, 5(2): 102-6, 1985.
- Plotkin, D.A., Mintz, J., Jarvik, L.F. Subjective memory complaints in geriatric depression. *American Journal of Psychiatry*, Sep, 142(9): 1103-1105, 1985.
- Sorenson, R.L., Gorsuch, R.L., Mintz, J. Moving targets: patients' changing complaints during psychotherapy. *Journal of Consulting and Clinical Psychology*, Feb, 53(1): 49-54, 1985.
- Van Putten, T., Marder, S.R., Mintz, J. Plasma haloperidol levels: clinical response and fancy mathematics. *Archives of General Psychiatry*, Aug, 42(8): 835-838, 1985.
- Albers, L.J., Doane, J.A., Mintz, J. Social competence and family environment: 15-year follow-up of disturbed adolescents. *Family Process*, Sept, 25: 379-389, 1986.
- Jenkins, J.H., Karno, M., de la Selva, A., Santana, F., Telles, C., Lopez, S., Mintz, J. Pharmacotherapy, expressed emotion and schizophrenic outcome among Mexican-Americans. *Psychopharmacology Bulletin*, 22 (3): 621-627, 1986.
- Marder, S.R., Hawes, E.M., Van Putten, T., Hubbard, J.W., McKay, G., Mintz, J., May, P.R.A., Midha, K.K. Fluphenazine plasma levels in patients receiving low and conventional doses of fluphenazine decanoate. *Psychopharmacology*, 88: 480-483, 1986.
- Marder, S.R., Hubbard, J.W., Van Putten, T., Hawes, E. M., McKay, G., Mintz, J., May, P.R.A., Midha, K.K. Plasma fluphenazine levels in patients receiving two doses of fluphenazine decanoate. *Psychopharmacology Bulletin*, 22(1): 264-266, 1986.

- Doane, J.A., Mintz, J. Communication deviance in adolescence and adulthood: a longitudinal study. *Psychiatry*, Vol. 50, Feb, 1987.
- Karno, M., Jenkins, J.H., de la Selva, A., Santana, F., Telles, C., Lopez, S., Mintz, J. Expressed emotion and schizophrenic outcome among Mexican-American families. *Journal of Nervous and Mental Disorders*, 175: 143-151, 1987.
- Marder, S.R., Van Putten, T., Mintz, J., Lebell, M., McKenzie, J., May, P.R.A. Low and conventional dose maintenance therapy with fluphenazine decanoate: two year outcome. *Archives of General Psychiatry*, 44: 518-521, 1987.
- Mintz, L.I., Liberman, R.P., Miklowitz, D.J., Mintz, J. Expressed emotion: a call partnership among relatives, patients and professionals. *Schizophrenia Bulletin*, 13, 227-235, 1987.
- Mintz, J., Mintz, L., Goldstein, M.J. A rejoinder to MacMillan et al. *British Journal of Psychiatry*, 151: 314-320, 1987.
- Van Putten, T., Marder, S.R., Mintz, J. The therapeutic index of haloperidol in newly admitted schizophrenic patients. *Psychopharmacology Bulletin*, 23: 201-207, 1987.
- Wilkens, J.N., Marder, S.R., Van Putten, T., Midha, K.K., Mintz, J., Setoda, D., May, P.R.A. Circulating prolactin predicts risk of exacerbation in patients on depot fluphenazine. *Psychopharmacology Bulletin*, 23: 522-525, 1987.
- Asarnow, R.F., Marder, S.R., Mintz, J., Van Putten, T., Zimmerman, K.E. The differential effect of low and conventional doses of fluphenazine on schizophrenic outpatients with good or poor information processing abilities. *Archives of General Psychiatry*, 45: 822-826, 1988.
- Miklowitz, D.J., Goldstein, M.J., Nuechterlein, K.H., Snyder, K.S., Mintz, J. Family factors and the course of bipolar affective disorder. *Archives of General Psychiatry*, 45: 225-231, 1988.
- Van Putten, T., Marder, S.R., Mintz, J., Poland, R.E. Haloperidol plasma levels and clinical response: A therapeutic window relationship. *Psychopharmacology Bulletin*, 23: 172-175, 1988.
- Hahlweg, K., Goldstein, M.J., Nuechterlein, K.H., Magana, A., Mintz, J., Doane, J.A., Miklowitz, D.J., Snyder, K.S. Expressed emotion and patient-relative interaction in families of recent onset schizophrenia. *Journal of Consulting and Clinical Psychology*, 57: 11-18, 1989.

- Marshall, B.D., Glynn, S., Midha, K., Hubbard, J., Bowen, L., Banzett, L., Mintz, J., Liberman, R. Adverse effects of fenfluramine in treatment refractory schizophrenia. *Journal of Clinical Psychopharmacology*, 9(2): 110-115, 1989.
- Mintz, L.I., Nuechterlein, K.H., Goldstein, M.J., Mintz, J., Snyder, K.S. The initial onset of schizophrenia and family expressed emotion: Some methodologic... considerations. *British Journal of Psychiatry*, 154: 212-217, 1989.
- Van Putten, T., Marder, S.R., Aravagiri, M., Chabert, N., Mintz, J. Plasma homovanillic acid as a predictor of response to fluphenazine treatment. *Psychopharmacology Bulletin*, 1: 89-91, 1989.
- Green, M.F., Nuechterlein, K.H., Ventura, J., Mintz, J. The temporal relationship between depressive and psychotic symptoms in recent-onset schizophrenia. *American Journal of Psychiatry*, 147: 179-182, 1990. Also published in Spanish: (1991). Relacion temporal entre los sintomas depresivos y psicoticos en la esquizofrenia de inicio reciente. *Psychiatry Digest*, 2: 19-20.
- Massel, H. K., Liberman, R. P., Mintz, J., Jacobs, H. E., Rush, T. V., Giannini, C. A., Zarate, R. Evaluating the capacity to work of the mentally ill. *Psychiatry* 53: 31-41, 1990.
- Sullivan, G., Marder, S.R., Liberman, R.P., Donahoe, C.P., Mintz, J. Social skills and relapse history in outpatient schizophrenics. *Psychiatry*, 53: 340-345, 1990.
- Van Putten, T., Marder, S.R., Mintz, J. A controlled dose comparison of haloperidol in newly admitted schizophrenic patients. *Archives of General Psychiatry*, 47: 754-758, 1990.
- Marder, S.R., Van Putten, T., Aravagiri, M., Hawes, E.M., Hubbard, J.W., McKay, G., Mintz, J., Midha, K.K. Fluphenazine plasma levels and clinical response. *Psychopharmacology Bulletin*, 26: 256-259, 1990.
- Crits-Christoph, P., Mintz, J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting and Clinical Psychology*, 59: 20-26, 1991.
- Marder, S.R., Midha, K.K., Van Putten, T., Aravagiri, M., Hawes, E.M., Hubbard, J.W., McKay, G., Mintz, J. Plasma levels of fluphenazine in patients receiving fluphenazine decanoate: Relationship to clinical response. *British Journal of Psychiatry*, 158: 658-665, 1991.

- Marder, S.R., Van Putten, T., Aravagiri, M., Hawes, E.M., Hubbard, J.W., McKay, G., Mintz, J., Midha, K.K. Therapeutic plasma levels of depot neuroleptics. *Psychiatry: A World Perspective*, (3)111-116, 1991.
- Marder, S.R., Mintz, J., Van Putten, T., Lebell, M., Wirshing, W.C., Johnston-Cronk, K. Early prediction of relapse in schizophrenia: An application of Receiver Operating Characteristic (ROC) methods. *Psychopharmacology Bulletin*, 27: 79-82, 1991.
- Mintz, J., Phipps, C. C., Arruda, M. J., Glynn, S. M., Schneider, N. G., & Jarvik, M. E. Combined use of alcohol and nicotine gum. *Addictive Behaviors*, 16: 1-10, 1991.
- Green, M.F., Mintz, J., Nuechterlein, K.H., Ventura, J. Postpsychotic Depression: Dr. Green and associates reply. *American Journal of Psychiatry*, 148: 546, 1991.
- Van Putten, T., Aravagiri, M., Marder, S.R., Wirshing, W.C., Mintz, J., Chabert, N. Plasma fluphenazine levels and clinical response in newly admitted schizophrenic patients. Abstracts of the III International Congress on Schizophrenia Research, Tucson, AZ, Apr 21-25, 1991. In *Schizophrenia Research*, 4(3): 295, 1991.
- Altman, E.S., Rea, M.M., Mintz, J., Miklowitz, D.J., Goldstein, M.J., Hwang, S.S. Prodromal symptoms and signs of bipolar relapse: A report based on prospectively collected data. *Psychiatry Research*, 41: 1-8, 1992.
- Rosenfarb, I., Mintz, J. Assessing social skill in role-play scenes: Is personal relevance relevant? *Behavioural Psychotherapy*, 20: 141-145, 1992.
- Dawson, M.E., Nuechterlein, K.H., Schell, A.M., Mintz, J. Concurrent and predictive electrodermal correlates of symptomatology in recent-onset schizophrenic patients. *Journal of Abnormal Psychology*, 101: 153-164, 1992.
- Mintz, J., Mintz, L.I., Arruda, M.J., Hwang, S. Treatments of depression and the functional capacity to work. *Archives of General Psychiatry*, 49: 761-768, 1992.
- Nuechterlein, K.H., Dawson, M.E., Citlin, M., Ventura, J., Goldstein, M.J., Snyder, K.S., Yee, C.M., Mintz, J. Developmental processes in schizophrenic disorders: Longitudinal studies of vulnerability and stress. *Schizophrenia Bulletin*, 18: 387-425, 1992.
- Nuechterlein, K.H., Snyder, K.S., Mintz, J. Paths to relapse: Possible transactional processes connecting patient illness onset, expressed emotion, and psychotic relapse. *British Journal of Psychiatry*, 161: 88-96, 1992.

Wirshing, W.C., Marder, S.R., Johnston-Cronk, K., Lebell, M., Mackenzie, J., Mintz, J., Eckman, T., Liberman, R.P. Management of risk of relapse in schizophrenia. Abstracts of the VI Biennial European Workshop on Schizophrenia, Badgastein, Austria, January 26-31, 1992. In *Schizophrenia Research*, 6: 107-108, 1992.

Chapters and monographs

Luborsky, L., Mintz, J. The contribution of P-technique to personality, psychotherapy and psychosomatic research. In R.M. Dreger (Ed.), *Multivariate Analysis: Essays in honor of Raymond B. Cattell*, Claitors Publishing Division: Baton Rouge, LA, 1972.

Woody, G., Mintz, J., O'Hare, K., O'Brien, C., Greenstein, R. Diazepam use by patients in a methadone program. *Proceedings of the NAS/NRC Committee on Problems of Drug Dependence*, Washington, DC, National Academy of Sciences, 114ff, 1975.

Greenstein, R., O'Brien, C., Mintz, J., Woody, G., Hanna, N. Clinical experience with naltrexone in a behavioral research study: An interim report. *NIDA Research Monograph Series*, 9: 141-149, 1976.

Mintz, J. The role of the therapist in assessing psychotherapy outcome. In A. Gurman and A. Razin (Eds.), *The therapist's handbook for effective psychotherapy*, London: Pergamon Press, 588-601, 1977.

Nace, E., O'Brien, C., Mintz, J. Drinking problems among Vietnam veterans. In F.A. Scixas (Ed.), *Currents in alcoholism*, New York: Grune & Stratton, 315-324, 1977.

O'Brien, C., Greenstein, R., Woody, G., Mintz, J. Naltrexone in combination with behavior therapy. In *Drug use and abuse: An international perspective*, *NIDA Monograph Series*, 1977.

Marcovici, M., O'Brien, C., Woody, G., Mintz, J. A comparative study of Levo-alpha acetylmethadol and methadone in the treatment of narcotic addiction. In E. Gottheil and A.T. McLellan (Eds.), *Addiction research and treatment: Converging trends*. London: Pergamon Press, 91-101, 1978.

Mintz, J. Tailoring psychotherapy outcomes to fit the individual case: A review. Report to NIMH, Psychotherapy and Behavioral Intervention Section, Clinical Research Branch, Contract No. PLD05668-77, 1978.

Nace, E., O'Brien, C., Mintz, J., Ream, N., Meyers, A. Adjustment among Vietnam veteran drug users two years post service. In C. Figley (Ed.), *Stress disorders among Vietnam veterans: Theory, research and treatment*. New York: Brunner-Mazel, 71-128, 1978.

- Woody, G., Mintz, J., Tennant, F., O'Brien, C., McLellan, A. Usefulness of propoxyphene napsylate for maintenance treatment of narcotic addiction. *NIDA Research Monograph Series*, 27: 240-246, 1979.
- Woody, G.E., Tennant, F.S., McLellan, A.T., O'Brien, C.P., Mintz, J. Lack of toxicity of high dose propoxyphene napsylate when used for maintenance treatment of addiction. *National Institute for Drug Abuse Research Monograph Series*, 27: 435-440, 1979.
- Mintz, J., Kiesler, D. Individualized measures of psychotherapy. In Kendall, P & Butcher, J. (Eds.), *Handbook of research methods in clinical psychology*, New York: Wiley, 1981.
- Jarvik, L., Mintz, J., Steuer, J., Gerner, R., Aldrich, J., Hammen, C., Linde, S., McCarley, T., Motoike, P. Rosen, R. Comparison of tricyclic antidepressants and group psychotherapies in geriatric depressed patients: An interim analysis. In Clayton, P.J. & Barrett, J.E. (Eds.) *Treatment of Depression: Old Controversies and New Approaches*, New York: Raven Press, 299-308, 1982.
- Marder, S., Van Putten, T., Mintz, J., Lebell, M., McKenzie, J., Faltico, G. Maintenance therapy in schizophrenia: New findings. In Kane, J. (Ed.) *Strategies for drug maintenance in schizophrenia*. Washington, DC: American Psychiatric Association Press, pages 32-49, 1984.
- Jarvik, L.F., Mintz, J. Treatment of depression in old age: What works? In, Awad, A.G., Durost, W.O. McCormick, H., Meier, M.R. (Eds.) *Disturbed behavior in the elderly*. New York: Pergamon Press, 1986.
- Marder, S.R., Mintz, J., Van Putten, T. Lebell, M., McKenzie, J. Prodromal symptoms as predictors of relapse. Lieberman, J.A. and Kane, J.M. (Eds.) *Clinical Insights: Predictors of Relapse in Schizophrenia*, American Psychiatric Press, Inc., Washington, D.C. Chapter 4. 47-57, 1986.
- Steuer, J.L., Mintz, J., Jarvik, L.F. Geriatric depression: Methodological issues in comparing pharmacology with group psychotherapy results. In, Miller, N & Cohen, G. (Eds.) *Psychodynamic research perspectives on development, psychopathology and treatment in later life*. New York: International Universities Press, Inc., 1986.
- Green M. F., Nuechterlein, K. H., Mintz, J., and Ventura J. The development of methods to assess the temporal relationship of depressive and psychotic symptoms in schizophrenia. In R. Williams and T. Dalby (Eds.) *Depression in Schizophrenics*. St. Louis: Plenum, 1989.

- Van Putten, T., Marder, S.R., Mintz, J., Poland, R.E. Haloperidol plasma levels and clinical response: A therapeutic window relationship. In S.C. Schulz & C.A. Tamminga (Eds.), *Schizophrenia: Scientific Progress*. New York: Oxford, 1989.
- Marder, S.R., Van Putten, T., Aravagiri, M., Hawes, E.M., Hubbard, J.W., McKay, G., Mintz, J., Midha, K.K. Therapeutic plasma levels of depot neuroleptics. In C.N. Stefanis, A.D. Rabavilas, C.R. Soldatos (Eds.), *Psychiatry: A World Perspective (Vol. 3)*. Amsterdam: Excerpta Medica, 111-116, 1990.
- Mintz J, Mintz LI, Phipps C. Treatments of mental disorders and the functional capacity to work. Chapter 13 in, R.P. Liberman (Ed.), *Handbook of Psychiatric Rehabilitation*, New York: Macmillan, 1991, 290-316.
- Nuechterlein KH, Dawson ME, Ventura J, Fogelson D, Citlin M, Mintz J. Testing vulnerability models: Stability of potential vulnerability indicators across clinical state. In H. Hafner & W. F. Gattaz (Eds.), *Search for the causes of schizophrenia*, Vol. II (pp. 177-191). Heidelberg: Springer-Verlag, 1991.
- Van Putten T, Marder SR, Wirshing WC, Chabert N, Mintz J, Aravagiri M. Plasma fluphenazine levels and clinical response in newly admitted schizophrenic patients. In S.C. Schultz, C.A. Tamminga (Eds.), *Schizophrenia: Scientific Progress*. New York: Oxford University Press, 1992.

Books

- Luborsky, L., Crits-Christoph, P., Mintz, J., Auerbach, A. *Who will benefit from psychotherapy? Predicting therapeutic outcomes*. Basic Books: New York, 1988.

Appendix B.

Vanderbilt's Technical Response to HCSCI Report CR93-002

Appendix B, page 1

This is a technical response to CR93-002 "Assessment of Two Data Collection Approaches for Fort Bragg Child/Adolescent Mental Health Demonstration Project Using Power Analysis" dated June 4, 1993 by the Army's Directorate of Health Care Studies and Clinical Investigation (HCSCI). The HCSCI report was received by Vanderbilt on July 8, 1993. This Appendix explains important areas of technical disagreement between the Vanderbilt research team and HCSCI report. Bold italic headings indicate quotes from the HCSCI report, a copy of which is included in Appendix E.

"the type of variable(s) used to measure 'improvement' between an average Demonstration and an average Comparison case was not defined." (p. 1)

Research team's response: Appendix C-4 of the HCSCI report lists the following Vanderbilt assumptions about variables used to define improvement: "The CBCL has a mean of 50 and SD 10 for normal children. Ours are in the mid 60's. . . . We assume that time and treatment make everyone average 0.3 SD better. The Demo provides an additional 0.25 sigma by fitting Tx [treatment] better to more children. The between-wave cross correlations are about $r = 0.50$ for adjacent waves, and about $r = 0.25$ for nonadjacent waves." In the HCSCI report's Appendix C-4, hypothetical mean scores for Waves 2 - 3 - 4 were listed for the Demonstration: 66, 63, 60. For the Comparison: 66, 64.5, 63."

These handwritten notes in the HCSCI report's Appendix C were prepared for face-to-face presentation to HCSCI's consultant, Dr. Wojcik, who did not keep her requested appointment at Vanderbilt. She did not reschedule. Without face-to-face meetings over calculations, it would be very difficult for scientists unfamiliar with the project (and inexperienced in mental health research) to grasp the complex analysis of the Ft. Bragg outcome study and the power analysis.

When Dr. Wojcik failed to appear for the arranged appointment at Vanderbilt, an eighteen-page summary of the power analysis was written by E.W. Lambert, Ph.D. This document, "Power analysis of the Ft. Bragg Evaluation Project: Technical details of a practical Monte Carlo power analysis [5/5/93]," was sent by FAX and Federal Express to L. Colonel T. E. Leonard at Ft. Sam Houston on 5/10/93. Vanderbilt received no response from the Army about this report.

When the HCSCI report of June 4 states that "the type of variable(s) used to measure 'improvement' between an average Demonstration and an average Comparison case was not defined" it also ignores the following material from Vanderbilt's May 5, 1993 summary:

For example, the Ft. Bragg Evaluation Project presents the following characteristics, all of which are important in determining the study's power:

- 1.[*] Two experimental groups were subjected to two forms of treatment (called "Demonstration" and "Comparison");
2. Clinical status was measured by continuous variables (such as the Achenbach CBCL total score) at three waves: admission (Wave 1), and 6- and 12-month follow-up (Waves 2 and 3);
3. Wave 1 scores had a mean of t-score 65. All scores have a standard deviation (SD) of about 10. Achenbach's norms suggest that normal children not in treatment have a mean of 50 and a SD of 10.

4. On the average, patients in both groups improve due to [time + treatment + regression to the mean]. This effect is of little interest in this study, since we are interested in difference between treatment methods across time, but the wave effect on all subjects had to be included in a complete data model. The effect size of [time + treatment + regression to the mean] was 0.30 SD.

5. We hypothesized an effect size in which all patients improved an average of 0.30 SD (from t-score = 65 to 62) and patients in the Demonstration improved another 0.25 SD (from 62 to 59.5) by Wave 3. This means that a Comparison child having a score of t-score 65 on Wave 1 intake would have a score of $65.0 - 3.0 = 62.0$ on Wave 3 one year later. A Demonstration child having a score of 65 on Wave 1 intake would have a score of $65.0 - 3.0 - 2.5 = 59.5$ on Wave 3 one year later. This effect size (2.5 points or 0.25 SD) was chosen, rather than a larger one, because 1) the Ft. Bragg Demonstration is a mental health system study, not a focused university-based study of a well-defined treatment vs. a well-defined nontreatment condition; and, 2) Many patients in both groups were evaluated on intake, did not return for treatment, but were evaluated in the study. These "nontreated" cases cannot be ignored when we study a mental health system, but such individuals dilute the larger effects in patients who receive regular treatment for a year or more.

6.[*] The number of subjects will be unbalanced, because larger numbers of subjects have been recruited in the Demonstration.

7. Correlations between Wave 1, Wave 2, and Wave 3 would be $r(1,2) = 0.50$ and $r(1,3) = 0.25$. These cross-wave correlations¹ occurred when a child's status at intake carries carry over somewhat to Wave 2. While the data are not yet in, the correlation between Waves (1, 3) is probably less than the correlation between Waves (1, 2) and between Waves (2, 3). This persistence, an autoregressive effect in which scores carry over time, can make ordinary least squares statistics show significance when effects are actually due to chance. Appendix A [of original report being quoted] shows cross-wave correlations for actual CBCL data taken from children (not computer-generated).

*Note: The two items marked with asterisks are Vanderbilt data assumptions used in the HCSCI power analysis; all the other data characteristics listed by the Vanderbilt researchers were ignored by HCSCI's t-test model. Ignoring assumption 2 (Wave1-Wave2-Wave3) is a particularly serious problem in the HCSCI analysis.

¹Actual data suggest that the adjacent wave correlation is around 0.5 or 0.6; the reduction to 0.25 on nonadjacent waves is an educated guess.

Appendix C of the HCSCI report shows data characteristics of the Monte Carlo simulation computed by Vanderbilt. Included are means, standard deviations, skews, kurtosis, quartiles, percentiles, ranges, modes, and other details for three waves for 80,000 model-generated subjects. Vanderbilt's cross-correlations and Demonstration-Comparison differences across three waves also appear in an appendix to the HCSCI report. If HCSCI did not understand these data, they should have completed a site visit to learn the details of Vanderbilt's power analysis.

Vanderbilt gave detailed descriptions of data assumptions. HCSCI's statement that *"the type of variable(s) used to measure 'improvement' between an average Demonstration and an average Comparison case was not defined"* is inaccurate, misleading, and unfair.

"The effect size (ES = 0.25) . . . should be used with caution." (p. 2)

We agree that caution is often beneficial in research. We also believe that effect size is "difficult to assess" and that "ES = 0.25 may or may not express the . . . actual variables measuring health outcome² [sic] in the Fort Bragg Evaluation Project." It is difficult to find the correct effect size before conducting a study; if we knew the effect size, we would not need to conduct the research. If the consultants were aware of a better meta-analysis on this issue, or some other superior way of knowing an effect size before the study is completed, they should have provided that information. However, we can only assume the HCSCI team could find no better estimate than $E.S. = 0.25$, or they would have suggested it.

". . . only a pilot test would give an answer as to the probable magnitude of the ES index." (p. 3)

This criticism by HCSCI reflects a basic misunderstanding of how researchers use power analysis to plan research. Power analysis is normally done when the study is designed, before data are available. The Vanderbilt research team believes that "To obtain the power associated with a study on treatment effectiveness, all one needs is some assumptions on the variance of the two treatment outcomes (in this study demonstration and control cases), the number of individuals in each group, the effect size and the level of significance. Power calculation does not require a 'peek' at the actual data." This statement was written by Dr. Kapadia on May 10, 1993 in her letter to Admiral Edward D. Martin³. Admiral Martin wrote: "I believe the Army has located an eminently qualified individual in Dr. Asha Kapadia . . ."

The Vanderbilt research team heartily recommends that the HCSCI group consider Dr. Kapadia's view of power analysis. The Vanderbilt researchers have tried to follow this view despite Health Services Command's (HSC) demands for client data it sought to use somehow in power analysis (see Appendix F). Dr. Kapadia's statement is the standard view held by power-analysis authors (e.g., Cohen, 1992; Lipsey, 1990). However,

²The Ft. Bragg Evaluation concerns mental health outcome.

³She said "I have now completed my review of the materials submitted to my on April 23, 1993, by Dr. Scott Optenberg." [Materials not seen by Vanderbilt until July 8, 1991.] A copy of Dr. Kapadia's letter is included in Appendix C.

on May 21, 1991, the Army's Contracting Officer, Leo M. Sleight⁴, wrote Dr. Behar demanding "all currently collected patient enrollment . . . all currently collected workload data . . . data resulting from tests . . . all summary data . . . all variable definitions" for the Army's power analysis. This demand for data on May 21 ignored Dr. Kapadia's good advice on May 10.

The Army's consultants seem to disagree with each other. The Vanderbilt research team agrees with Dr. Kapadia, who expresses the standard view of power analysis based on assumptions about the data. We disagree with HCSCI's statement: *"Since the Monte Carlo technique presented in Appendix C does not involve actual data, the results from this method may be entirely misleading and not accurate."*

When HSC demanded client data for power analysis in order to decide when to stop gathering data, Vanderbilt was greatly concerned about damage to the study's final results, especially to a possible crippling of the final report of outcome. According to the Pharmaceutical Manufacturers Association's Biostatistics and Medical Ad Hoc Committee on Interim Analysis⁵, "If the trial is terminated early due to an unscheduled interim analysis, there are no established statistical theories to compensate for the effect of such procedures on the α level" (p. 163). In other words, the whole outcome analysis could be clouded if the data were analyzed and the decision to stop adding cases were made based on the results.

"Power analysis of Short and Long-Term Plans⁶" (pp. 3-7)

HCSCI's power analysis analyzes the Ft. Bragg Evaluation with a one-tailed t-test on Wave 3. The problem here does not concern the statistical details of the HCSCI report's power estimates; the table "lookup" was done correctly. Using a t-test on the Wave 3 post-test to evaluate the results of the Ft. Bragg Evaluation, however, reveals fundamental ignorance of the study's purpose and its experimental design. It is a fatal flaw in HCSCI's proposed power estimates. We discuss the implications of this choice below in terms of the Evaluation's design and data structure.

In the Ft. Bragg evaluation, the Army funded three waves of data collection for two sites: Demonstration (Ft. Bragg) and Comparison (Ft. Stewart and Ft. Campbell). Thousands of variables are being collected across three waves [Wave 1 - intake, Wave 2 -6 months, Wave 3 -12 months] to describe the child's problems, the family environment, school performance, etc. in exhaustive detail. Vanderbilt's power analysis was based on a repeated measures ANOVA or MANOVA of two groups by three waves:

Demonstration	Wave 1	Wave 2	Wave 3
Comparison	Wave 1	Wave 2	Wave 3

⁴Mr. Sleight's letter in Appendix F of the present report.

⁵This is an official panel representing researchers and manufacturers with serious concerns about terminating expensive clinical trials.

⁶Note by Vanderbilt: "Short-term" refers to the Army plan, "long term" to Vanderbilt's.

Appendix B, page 5

This analysis is designed to measure change across the three waves. We want to know how children change after intake. The time by treatment interaction used in our power analysis answers the following question: Does the average change after intake, and across time, by Demonstration children differ from the average change of Comparison children, and, if so, is the Demonstration better or worse?

HCSCI's power analysis uses the following design:

Demonstration	xxxxx	xxxxx	Wave 3
Comparison	xxxxx	xxxxx	Wave 3

According to the HCSCI report, we should to a t-test at Wave 3. In this design, "xxxxx" stands for the data the Army paid for but HCSCI ignored (i.e., proposed never to analyze) by proposing its t-test. The HCSCI t-test answers the question: Which group had a higher average score on Wave 3 (end of treatment)? HCSCI's approach ignores the issue of change during treatment. It also ignores pretest site differences.

The "posttest only" design suggested by HCSCI is described in two classic books on research design (Campbell and Stanley, 1963; Cook and Campbell, 1979). Campbell and Stanley commend the posttest-only control group experimental design for randomized experiments, viz. studies in which patients are randomly assigned to treatment.⁷ However, the Evaluation is not a randomized experiment. The Ft. Bragg Evaluation is a nonequivalent group design with no guarantee that children at Ft. Bragg are the same at intake as children at Campbell or Stewart.

Posttest only studies without random assignment are discussed by Cook and Campbell (1979) under the heading "Three designs that often do not permit reasonable causal inferences." Cook and Campbell (1979, p. 95) explain, "Its most obvious flaw is the absence of pretests, which leads to the possibility that any posttest differences between groups can be attributed either to a treatment effect or to selection differences⁸ between the different groups. The plausibility of selection differences in research with nonequivalent groups usually renders the design uninterpretable."

In other words, the significant differences in a posttest-only t-test without random assignment may be due to treatment effects or they may be due to pre-existing site differences. The t-test inextricably confounds the two effects. It doesn't tell us what we want to know: Did children change for the better in the Demonstration?

⁷The Ft. Bragg Evaluation is not an experiment in which clients are assigned randomly to treatment. If the Army team mistakenly believed that children were assigned at random to clinics hundreds of miles apart, then their t-test would be statistically correct. Gathering Wave 1 and Wave 2 data would then be unnecessary.

⁸We know already that there are significant differences on clinical summary scores, such as CBCL scores, between the sites. If the means differ on Wave 3, a t-test on Wave 3 may be different because children got better during treatment, or merely that one group started out better. The Ft. Bragg Evaluation concerns the effect of treatment (i.e., changes after intake during treatment).

In addition to the fundamental misunderstanding above, the HCSCI report's choice of one-tailed tests reflects technical error: ". . . testing the null hypothesis that $m_d = m_c$ at $\alpha_1 = 0.05$ (one-tailed test) (Table 2.3.2 from Cohen, 1988)." (p. 3)

The HCSCI report specifies a two-tailed hypothesis ($m_d = m_c$) and then performs a one-tailed test⁹.

The null hypothesis $m_d = m_c$ refers to a two tailed test (H_0 : means are equal) not a one-tailed hypothesis (H_0 : $M_a > M_b$). According to Cohen (1988, p. 27) ". . . for one-tailed tests . . . the alternative hypothesis specifies that $m_b > m_a$."

The difference between one-tailed and two-tailed tests are explained below.

Two-tailed tests. Treatment researchers normally test two-tailed nondirectional hypotheses, such as "the null hypothesis that $m_d = m_c$ " [Quoted from HCSCI's description of their method.] This null hypothesis states that if the difference in either direction is large enough, we have results that are not due to chance. In HCSCI's notation $m_c = m_d$ [null hypotheses, mean of comparison equals mean of demonstration]. Demonstration children may fare better than Comparison children, or Comparison children may fare better than Demonstration children; in either case we have statistically significant results, and a lesson worth reporting.

One-tailed tests. In a one-tailed directional test, we might hypothesize $m_c < m_d$ (the mean of Comparison is smaller than the mean of Demonstration), or else we might say $m_c > m_d$ (the mean of Demonstration is smaller than the mean of Comparison), and do a one-tailed test. The reason why published mental health research seldom uses directional tests follows: If the researcher hypothesizes directionally that Demonstration children do better and then finds that Demonstration children do dramatically worse, the researcher cannot report that result. This absurdity makes directional hypotheses very rare in treatment research.

Using Cohen's one-tailed power table gives an estimate requiring many fewer subjects than a two-tailed test. For example, with an effect size of 0.20, one-tailed power of 80% requires $n \approx 300$ in each sample. Two-tailed requires $n \approx 400$. The misguided choice of a one-tailed t-test makes the analysis appear more powerful. This added power supports the view that the Army should invest less money in research, but it is not technically correct. To be technically correct, the one-tailed directional hypothesis must be stated in advance. For example, "We hypothesize that the Demonstration is better. If the Comparison is better, we will not report the result." The null hypothesis stated by the HCSCI report ($m_d = m_c$) is the correct two-tailed null hypothesis, but they then used one-tailed table "lookups".

"Assessment of the simulation method." (pp. 7-9)

The second half of the HCSCI report is titled "Assessment of the simulation method." Before responding to the opinions stated by HCSCI, a brief summary of computer simulation is needed. This summary was taken from the document, "Power analysis of the Ft. Bragg

⁹A one-tailed test requires fewer clients than a two-tailed test.

Evaluation Project: Technical details of a practical Monte Carlo power analysis [5/5/93]," which was sent to L. Colonel T. E. Leonard at Ft. Sam Houston on 5/10/93. A copy of the document is included in Appendix G. To date, the Army has not responded to that document.

Summary of Monte Carlo power analysis. Simple projects can use simple techniques available in standard texts to calculate the power of their experimental designs. The most common method of power analysis is to find one's design in a power analysis text, such as Cohen (1988) or Lipsey (1990), choose the appropriate effect size, and look up the power from power curves or power tables. Computer programs, such as BMDP Solo (Hintz, 1992) automate the process and draw power curves to fit a particular situation. While this "look-up" approach works well for common problems, such as the two group t-test, advanced large-scale evaluations often have features not found in standard look-up tables.

By analyzing scores with known characteristics, including an experimental effect, a large number of times and counting how many times the effect was detected or missed, estimates of power were observed as simple counts. For example, if in 150 analyses, each with 820 subjects, the experimental effect is detected 79% of the time, one would conclude that the estimated power is 79% at $N = 820$ subjects, and that the β -error rate is 21% (i.e., 21% of the time an effect exists and we fail to detect it).

In other words, a Monte Carlo simulation generates data according to assumptions, and then analyzes it many times using whatever analytic technique is needed.

Review of "Assessment of simulation method". This section in the HCSCI report states the following opinions:

1. Simulation is not a substitute for knowledge (p. 7);
2. The use of simulation requires complete information (p. 8);
3. [simulation] . . . does not involve actual data¹⁰ (p. 9);
4. [simulation] should only be utilized when direct data analysis cannot (p. 9);
5. only one variable was used (p. 9);
6. without the use of actual data, the effect size value . . . was used to calculate the power in this report. (p. 9)

These nonspecific criticisms apply to any of simulation in which hypothetical data is used because real data are unavailable. The HCSCI report's simulation of the outcome analysis of the Ft. Bragg Evaluation with a table "lookup" from Cohen's book -- doing the power analysis based on assumptions -- has all of these same problems. We agree that analyzing the actual data that results from the Demonstration will be more enlightening than analysis based on assumptions, but we cannot analyze actual data until after they are collected. Obviously, then it will be too late to plan how much data to gather.

In criticizing the Monte Carlo simulation as a bad method, the HCSCI report confuses methods with results. HCSCI's nonspecific denigration of the Monte Carlo simulation

¹⁰According to the Army's consultant, Dr. Kapadia, power analysis should not use actual data. It should use assumptions about the data to be gathered.

implies that the Army result is right because it came out of a book while the Vanderbilt result is wrong because it came out of a computer.

To determine specifically whether the Monte Carlo simulation gave different results, we replicated the HCSCI report's two-group t-test power analysis with five methods: the Vanderbilt Monte Carlo, and four published approaches, two from books (Cohen, 1988; Lipsey, 1990), the other two with software (Dupont and Plummer, 1990; Hintz, 1991). The results appear in Figure 1; the published methods showed very close agreement with each other. The Monte Carlo power analysis (heavy line with stars), while not as smooth¹¹ as equation-based simulations, gave results consistent with the other curves when applied to the posttest only t-test¹². These power analyses were done with the same minimal assumptions as the HCSCI report, viz. a two-group posttest only t-test¹³ with one group's mean 0.25 SD above the other's. Cohen's "lookup" tables had no effect size $E.S. = 0.25$, so $E.S. = 0.20$ and 0.30 were both plotted. Cohen's $E.S. = 0.25$ would appear between $E.S. = 0.20$ and 0.30 .

If we compare the merits of the two methods, customary ones (i.e., "lookup" tables and commercial software) are quick and easy to use, but they are not available for every analysis. Monte Carlo simulation requires more work, but it fits any data model that can be analyzed on a computer. Both methods have a legitimate place; either method can be applied in error or abused.

"Conclusion" (p. 9)

In a section titled "Conclusion," the HCSCI report actually states three conclusions:

1. ***"The power values for the directional tests computed in this [HCSCI] study and the values given in the [Vanderbilt] proposal are significantly different."*** The research team agrees with this conclusion. Vanderbilt used a Wave 1 - Wave 2 - Wave 3 repeated measures ANOVA and MANOVA, whereas the HCSCI report used a posttest-only t-test. The more complex design reveals the results of treatment, but it is less powerful. The t-test is more powerful (requires fewer subjects), but it throws away two-thirds of the data and cannot tell us whether kids improved more in the Demonstration.
2. ***"Secondly because the standardized effect size is a computed variable it can be modified . . . Variance can be reduced . . . unnecessary dichotomization causes a loss of power."*** While the HCSCI report seems aware that normed rating scales, such as the CBCL, are used in the study¹⁴, they do not seem aware that these normed instruments produce continuous scores, not dichotomies, and that normed instruments must be scored according to their author's instructions. Their second conclusion is analogous to telling psychologists to score an IQ test in novel ways so that people no longer differ in intelligence.

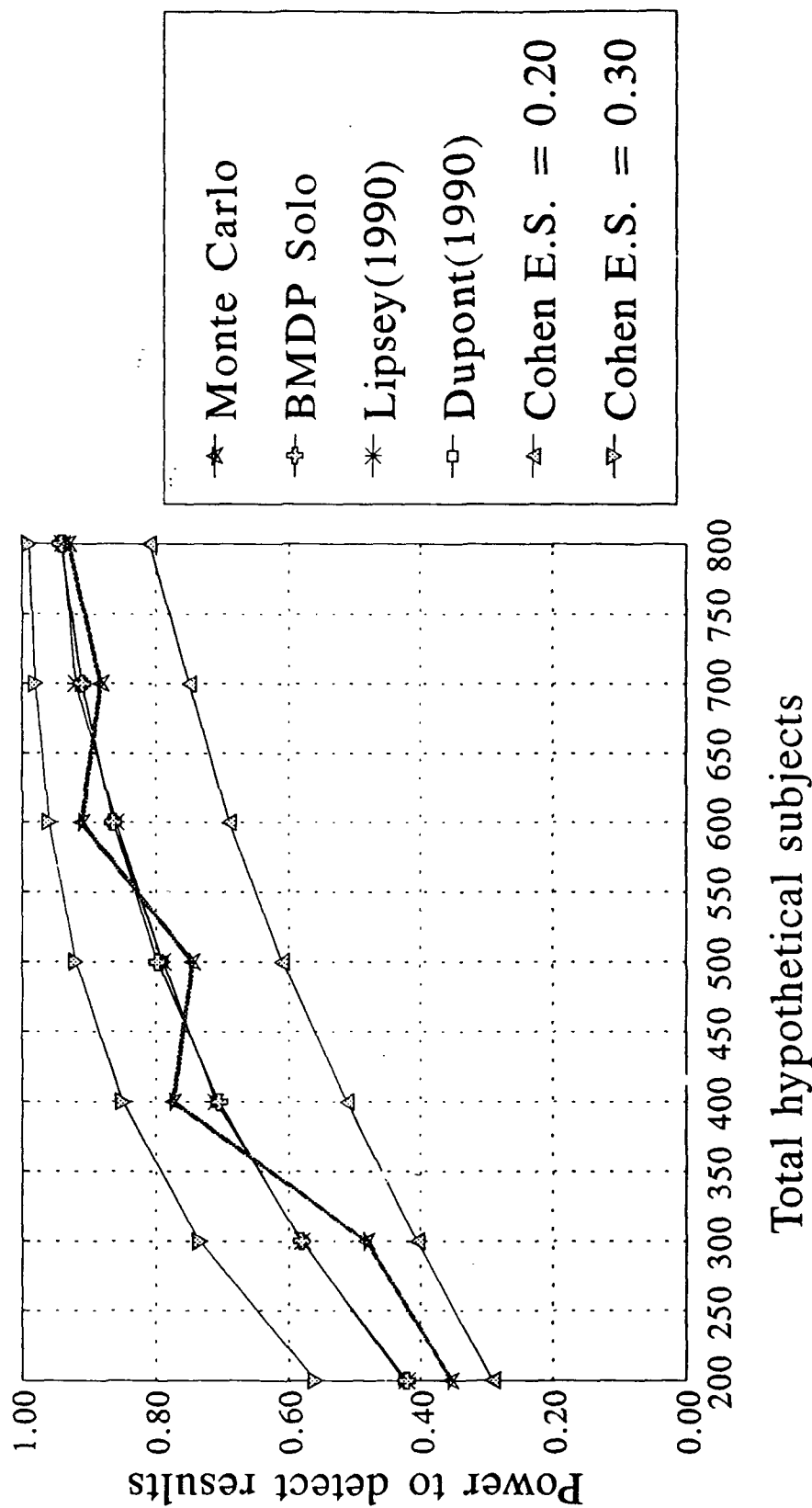
¹¹By running the simulation more hours on a desktop computer, the Monte Carlo could achieve any degree of precision and smoothness needed.

¹²The Monte Carlo simulation is unnecessary for a t-test, since we can simply look up the power for the t-test; the simulation is needed to model the list of data-based assumptions that the Ft. Bragg outcome study requires.

¹³Two-tailed tests were used with all methods.

¹⁴HCSCI Report CR93-002, p. 2.

Fig. 1. Monte Carlo Power Estimates
Compared with Four Published Methods
(Effect size = 0.25 S.D.)



Notes:

Post-test only t-test or ANOVA with equal-N groups.

This design confounds pretest differences with results of treatment.

3. ***"Finally . . . a more accurate estimate of the Fort Bragg Evaluation Project effect size is achieved when actual data is utilized . . ."***

The research team follows Cohen, Lipsey, Dupont, Hintz, and Kapadia: Power analysis is calculated based on assumptions before data are gathered. By the time a project has enough data to estimate its effect size accurately, it has completed the study.

References

- Campbell, D.T., & Stanley, J.C. (1979). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, N.J.: Erlbaum.
- Cohen, J. (1992) A power primer. *Psychological Bulletin*, 112(1), 155-159.
- Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand-McNally.
- Dupont, W.D. and Plummer, W.D. (1990). Power and sample size calculations: A review and computer program. *Controlled Clinical Trials*, 11, 116-128.
- Hintz, J. (1992). *Solo Statistical System Power Analysis*. Los Angeles, CA: BMDP Statistical Software.
- Lipsey, M.W. (1990). *Design Sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage Publications.
- Pharmaceutical Manufacturer's Association Biostatistics and Medical Ad Hoc Committee on Interim Analysis (1993). Interim analysis in the pharmaceutical industry. *Controlled Clinical Trials*, 14, 160-173.

Appendix C.
Letter from Dr. Kapadia,
Consultant to Health Services Command

DANIEL K. INOUE
HAWAII

APPROPRIATIONS
Chairman, Subcommittee on Defense

COMMERCE, SCIENCE AND TRANSPORTATION
Chairman, Subcommittee on Communications

Chairman, SELECT COMMITTEE ON INDIAN
AFFAIRS

Chairman, DEMOCRATIC STEERING COMMITTEE

Member, COMMITTEE ON RULES AND
ADMINISTRATION

United States Senate

SUITE 722, HART SENATE OFFICE BUILDING
WASHINGTON, DC 20510-1102
(202) 224-3934
FAX (202) 224-6747

PRINCE KUHIO FEDERAL BUILDING
ROOM 7325, 300 ALA MOANA BOULEV.
HONOLULU, HI 96850-4975
(808) 541-2542
FAX (808) 541-2549

101 AUPUNI STREET, NO. 205
HILO, HI 96720
(808) 935-0844
FAX (808) 961-5163

June 14, 1993

Dr. Leonard Bickman
Peabody College
Vanderbilt University
P.O. Box 163
Nashville, Tennessee 37203

Dear Dr. Bickman:

I am writing to share with you a copy of a report I recently received from Dr. Edward Martin, Acting Assistant Secretary of Defense for Health Affairs, in response to my request that he personally review the evaluation component of the U.S. Army Mental Health Demonstration Project at Fort Bragg, North Carolina.

It is my understanding that a final determination will soon be made by Dr. Martin, and that he would be pleased to meet to discuss this directly with you at the appropriate time.

Aloha,



DANIEL K. INOUE
United States Senator

DKI:phdt
Enclosure



THE ASSISTANT SECRETARY OF DEFENSE

WASHINGTON, D. C. 20301-1200

JUN 03 1993

HEALTH AFFAIRS

Patrick H. DeLeon, Ph.D.
Administrative Assistant
Office of Senator Daniel K. Inouye
United States Senate
Washington, DC 20510-1102

Dear Dr. ^{Pat}DeLeon:

Thank you for your letter of May 17, 1993, regarding the evaluation component of the Army's Mental Health Demonstration Project at Fort Bragg, North Carolina. We continue to agree that this is an important component of the project, and it needs to be conducted in a manner designed to produce the most useful and credible results.

Because of my concern that this be handled appropriately, and because the evaluator, Vanderbilt University, is under contract to the State of North Carolina and not to either the Army's Health Services Command or Surgeon General, I directed the Army to engage the services of someone with impeccable credentials to assist in resolving in differences in opinion and in helping both the Army and this office in evaluating the report from Vanderbilt. This seems particularly prudent since we are separated from the evaluator by the contractor whose results are being scrutinized.

I believe the Army has located an eminently qualified individual in Dr. Asha Kapadia at the University of Texas at Houston. Her curriculum vitae is enclosed for your information. Dr. Kapadia has also rendered an initial report on the main question at issue, i.e., the extension and increase in funding of the evaluation contract. This is also enclosed.

As you probably know, the National Institute on Mental Health is also funding Vanderbilt University for an evaluation of this demonstration. We anticipate that the results of that study will complement nicely the information garnered through the study funded by the Army through the State of North Carolina.

We are pursuing this matter with the utmost concern that we utilize the experience gained in a manner which recognizes the positive aspects of the demonstration at the same time it

identifies any lessons which need to be learned and incorporated in our plans and activities for future delivery of mental health services to our beneficiaries.

I appreciate your informed concern and commitment regarding health issues in the Military Health Services System. Your continued support is important to us.

Sincerely,

A handwritten signature in dark ink, reading "Edward D. Martin". The signature is written in a cursive style with a prominent initial "E".

Edward D. Martin, M.D.
Acting Assistant Secretary of Defense

Enclosures:
As stated

The University of Texas
Health Science Center at Houston



SCHOOL OF PUBLIC HEALTH
Health Services Organization

1200 Herman Pressler
P.O. Box 20186
Houston, Texas 77225
(713) 792-4372
(713) 792-4471

May 10, 1993

Edward D. Martin, M.D.
Assistant Secretary of Defense (Health Affairs)
The Pentagon, Washington DC

Dear Dr. Martin:

I have now completed my review of the materials submitted to me on April 23, 1993, by Dr. Scott Optenberg.

In the absence of information on several key factors relevant to the successful execution of a project of this magnitude, it is indeed impossible to conduct an objective evaluation of all the claims of the investigators for the Fort Bragg Demonstration project. I will therefore limit my comments to the power analysis performed by the Army Statisticians in an in house effort to determine if the demonstration project should be extended.

The investigators at Fort Bragg are interested in detecting a standardized difference of .25 between the experimental and control subjects for the short term plan. They anticipate 299 demonstration and 150 control cases at wave 3. As demonstrated by the detailed power analysis developed for this purpose by Dr. Optenberg's group, no matter what assumptions are made on the variances of the two populations, the minimum power that may be attained at wave 3 of the analysis is about 81%. The derivation of the power analysis is based on the theoretical developments presented in Cohen's (1988) book which is regarded as the basic text on power analysis in behavioral sciences.

Similarly, using the anticipated number of cases at the end of (wave 3) the long term plan, (i.e. 426 demonstration and 361 comparison cases) a power of at least 90% will be obtained.

In the Fort Bragg demonstration project, a power of .80 for detection of a relatively small difference (i.e. .25 SD) in improvement between subjects in the experimental and control groups is very impressive considering that most research studies in social sciences are under powered (power <.80) for detecting anything but large differences (Lipsey 1990). Thus, the short-term plan is more than sufficient to meet the objectives of this demonstration project.

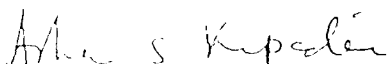
The investigators justification for a long term plan is based on the argument that if only the short term plan were to be carried out, the likelihood of detecting a statistical significance in the presence of a treatment effect would be 50%. This claim has not been demonstrated mathematically by the investigators and as shown by the power analysis performed by the army statisticians using appropriate statistical procedures is in serious error.

On reviewing the documentation dated April 30, 1993 from Vanderbilt University (received by me on 5/8/93), some inconsistency in the claims of the Investigators/Evaluators of the demonstration project is apparent. On page 4 of the above document, they state that the project has been losing about 15% of the subjects per wave. Using this attrition rate the 1065 wave 1 cases (demonstration plus control) should result in 1065 (.85)(.85) or 769 cases. Yet under data collection assumptions the 1065 wave 1 cases will result in only 449 (299 demonstration and 150 control) cases. Therefore, the statistical power under the proposed short-term plan may be even higher than 81%.

Furthermore, investigators have repeatedly mentioned not wanting to "peek prematurely" at the real data for fear of "ruining the chance to use standard statistical estimates in the way that they were designed". To obtain the power associated with a study on treatment effectiveness, all one needs is some assumption on the variance of the two treatment outcomes (in this study demonstration and control cases), the number of individuals in each group, the effect size and the level of significance. Power calculation does not require a "peek" at the actual data. Hence the use of Monte Carlo simulation to estimate the power of the study is unnecessary and irrelevant.

If I may be of further help, please feel free to call me at (713) 792-4472.

Sincerely,


Asha S. Kapadia
Professor and Convener
of Biometry

ASK:rf

Appendix D.
Vanderbilt's Technical Response to Dr. Kapadia's Letter

Apparently Dr. Kapadia was retained by HCSCI to review Vanderbilt's analysis and its own analysis. Vanderbilt does not know what information was provided to her by the Army. Our only contact was a call from Dr. Kapadia's office by Dr. Optenberg indicating that a report had been submitted by his group. However, he told Dr. Bickman that he could not discuss the report, nor could he reveal its conclusions to Vanderbilt. Neither of the consultants retained by HCSCI has ever been in contact with Vanderbilt, nor has Dr. Optenberg had any discussions of substance with Vanderbilt about the power analysis (included in Appendix G). Vanderbilt provides below a review of Dr. Kapadia's letter to Dr. Martin in an attempt to clarify misconceptions presented in that letter.

Support of Army statisticians. Dr. Kapadia's first six paragraphs support " . . . army statisticians using the appropriate statistical procedures . . ." However, she admits that her conclusion was "In the absence of information on several key factors relevant to the successful execution of a project of this magnitude" and that "it is impossible to conduct an objective evaluation of all the claims of the investigators . . . [and she] will therefore limit my comments to the power analysis performed by the Army Statisticians in an in house effort"

Her report does not say whether she reviewed the appropriateness of the t-test of means on Wave 3, nor whether the design and purpose of the Ft. Bragg Evaluation were explained to her by HCSCI. She did not comment on the following issues: Is a Wave 3 t-test the analysis she would use to determine the results of the Ft. Bragg project? Is a Wave 3 t-test what the investigators should deliver to the Army in the final report?

Given that a t-test on means would be appropriate, the Vanderbilt investigators agree with Dr. Kapadia that the Army statisticians' power analysis is correct (i.e., that they looked up the power accurately from Cohen's book). However, the t-test on Wave 3 confounds the effect of treatment with status at intake. An analysis of Wave 3 only would not produce any interpretable result for the Army, Congress, or the professions. The HCSCI report gives the correct power estimates for the wrong analysis.

"The Investigators' inconsistent claims". In paragraph seven, Dr. Kapadia believes she discovered "some inconsistency in the claims of the Investigators [Vanderbilt]." She reasoned that 1065 intakes with 15% attrition should produce $1065 * 0.85 * 0.85 = 769$ cases on Wave 3.

However, in the very paragraph she cites, the Vanderbilt investigators stated that "The short-term plan stops recruitment (Wave 1) at all sites on June 30, 1993 . . . and stops all data collection for Waves 2 and 3 on September 30, 1993." Perhaps Dr. Kapadia did not know that Wave 3 data are collected one year after Wave 1 intake. Counting back one year from the end of Wave 3 data collection (September 30, 1993), participants whose Wave 1 data are collected after September 30, 1992 (under the Army's short term plan) would not have the opportunity to provide Wave 3 data within the short-term data collection window. Those with Wave 1 intakes during the last nine months of the Army's proposed data collection all would be dropped without providing outcome (Wave 3) data. Under the long term plan, data collection halts one year after the last intake.

Appendix D, page 2

This oversight by Dr. Kapadia suggests that the Army did not, or perhaps could not, inform her fully and accurately about the nature of the Ft. Bragg Evaluation; otherwise she would have known that data collection lasted one year per subject, and that the short term plan dropped unfinished subjects above and beyond the 15% attrition estimate. In her letter, she seems to answer the question: Was the Army's power estimate for their analysis correct, not the more fundamental question: Was the Army's t-test on Wave 3 the appropriate statistic?

If there is inconsistency, it is with the Army's short-term plan. Why should the Army pay for intakes for nine months and not bother to gather Wave 3 data to determine the outcome?

Not "peeking" at the data. In paragraph eight, Dr. Kapadia appropriately points out that "power calculation does not require a 'peek' at the data." Indeed, Vanderbilt investigators repeatedly tried to convince the Army that analyzing incomplete client data would not help in power analysis. However, judging by HSCSI's report and HSC's request for data (see Appendix F), Vanderbilt failed to make this point. Perhaps the Army group will accept this basic orientation to power analysis coming from Dr. Kapadia, or from Cohen (1992).

Reference

Cohen, J. (1992) A power primer. *Psychological Bulletin*, 112(1), 155-159.

Appendix E.
HCSCI Report CR93-002



North Carolina Department of Human Resources
Division of Mental Health, Developmental Disabilities
and Substance Abuse Services

325 North Salisbury Street • Raleigh, North Carolina 27603 • Courier # 56-20-24

James B. Hunt, Jr., Governor
C. Robin Britt, Sr., Secretary

Michael S. Pedneau, Director
(919) 733-7011

July 8, 1993

Dr. Leonard Bickman
Vanderbilt University
Institute for Public Policy Studies
Box 7701 Station B
Nashville, Tennessee 37203

RE: DADA10-89-C-0013; Fort Bragg Child/Adolescent Mental Health
Demonstration Project; Technical Report of Power Analysis

Dear Len:

Enclosed is the report, "ASSESSMENT OF TWO DATA COLLECTION
APPROACHES FOR FORT BRAGG CHILD/ADOLESCENT MENTAL HEALTH
DEMONSTRATION PROJECT USING POWER ANALYSIS", that was prepared by the
Directorate of Health Care Studies And Clinical Investigation, United States Army Medical
Department Center and School, Fort Sam Houston, Texas.

I ask that you review and respond to the report as soon as possible. If you have
questions concerning the report, call me at (919) 733-0598.

Sincerely,

A handwritten signature in cursive script, appearing to read "Lenore", written over a horizontal line.

Lenore Behar, Ph.D.
Head, Child and Family Services Branch

Enclosure



DEPARTMENT OF THE ARMY
HEADQUARTERS, UNITED STATES ARMY HEALTH SERVICES COMMAND
FORT SAM HOUSTON, TEXAS 78234-6000



REPLY TO
ATTENTION OF

July 7, 1993

Central Contracting Office

SUBJECT: Contract DADA10-89-C-0013, Fort Bragg Mental
Health Demonstration Project; Technical Report on
Vanderbilt

Dr. Lenore Behar
North Carolina Department of Human Resources
Division of Mental Health, Developmental
Disabilities and Substance Abuse Services
325 North Salisbury Street
Raleigh, North Carolina 27603

Dear Dr. Behar:

This is in response to your letter dated June 8,
1993, requesting a copy of our independent analysis on
your proposed short term plan for the Vanderbilt
Evaluation Project.

We are sending you this report via Federal Express.
Request that you telephone to confirm receipt. Request
that you review the report and respond within 3 working
days from date of receipt.

If you have any further questions please contact
Ms. Joyce Nadeau, Central Contracting Office,
(210) 221-9453/0179.

Sincerely,

Leo M. Sleight
Leo M. Sleight
Contracting Officer, Central
Contracting Office

Agreed to ASH contract
lay

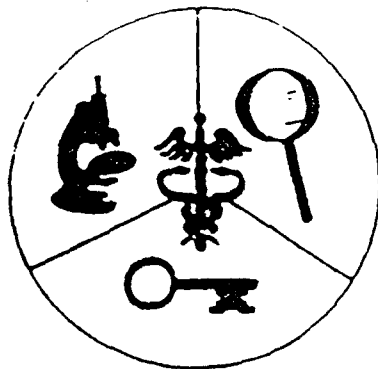
Enclosure

Copies Furnished:

Mr. James Newman, Contracting Officer's Representative,
Womack Army Medical Center, Fort Bragg,
North Carolina 28307-5000 (without report)
Lieutenant Colonel Leonard, Headquarters, U.S. Army
Health Services Command, Attention: HSCL-M,
Fort Sam Houston, Texas 78234-6000

JUL 1993

RECEIVED BY
[illegible]

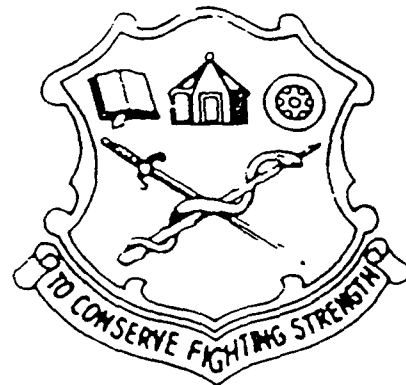


DIRECTORATE OF
HEALTH CARE STUDIES
AND CLINICAL INVESTIGATION

ASSESSMENT OF TWO DATA COLLECTION
APPROACHES FOR FORT BRAGG CHILD/ADOLESCENT
MENTAL HEALTH DEMONSTRATION PROJECT
USING POWER ANALYSIS

CR 93-002
PART I - FINAL REPORT
(REVISED)

JUNE 1993



UNITED STATES ARMY
MEDICAL DEPARTMENT CENTER AND SCHOOL
FORT SAM HOUSTON, TEXAS 78234-6100

JUL 1993

NOTICE

The findings in this report are
not to be construed as an official
Department of Defense position
unless so designated by other
authorized documents.

* * * * *

Regular users of services of the Defense Technical Information Center
(per DoD Instruction 5200.21) may purchase copies directly from the
following:

Defense Technical Information Center (DTIC)
ATTN: DTIC-DDR
Cameron Station
Alexandria, VA 22304-6145

Telephones: DSN 284-7633, 4, or 5
Commercial (703) 274-7633, 4, or 5

All other requests for reports will be directed to the following:

U.S. Department of Commerce
National Technical Information Service (NTIS)
5285 Port Royal Road
Springfield, VA 22161

Telephone: Commercial (703) 487-4600

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified			1b. RESTRICTIVE MARKINGS	
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION/AVAILABILITY OF REPORT Distribution Unlimited; Public Use Authorized.	
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE				
4. PERFORMING ORGANIZATION REPORT NUMBER(S) CR93-002 Part I - Final Report (Revised)			5. MONITORING ORGANIZATION REPORT NUMBER(S)	
6a. NAME OF PERFORMING ORGANIZATION Dir. Health Care Studies and Clinical Investigation		6b. OFFICE SYMBOL (If applicable) HSHN-A		7a. NAME OF MONITORING ORGANIZATION DASG
6c. ADDRESS (City, State, and ZIP Code) Bldg 2268 Fort Sam Houston, TX 78234-6000			7b. ADDRESS (City, State, and ZIP Code) Pentagon Washington, D.C. 20301	
8a. NAME OF FUNDING/SPONSORING ORGANIZATION HQ HSC		8b. OFFICE SYMBOL (If applicable)		9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) HQ HSC Fort Sam Houston, TX 78234-6100			10. SOURCE OF FUNDING NUMBERS	
			PROGRAM ELEMENT NO.	PROJECT NO.
			TASK NO.	WORK UNIT ACCESSION NO.
11. TITLE (Include Security Classification) (U) Assessment of Two Data Collection Approaches for Fort Bragg Child/Adolescent Mental Health Demonstration Project Using Power Analysis				
12. PERSONAL AUTHOR(S) Barbara E. Wojcik, Catherine R. Stein, M.S., Dr. Scott A. Optenberg				
13a. TYPE OF REPORT Final		13b. TIME COVERED FROM Apr 93 to May 93		14. DATE OF REPORT (Year, Month, Day) 1993 JUN 04
				15. PAGE COUNT 35
16. SUPPLEMENTARY NOTATION This is a report to the Assistant Secretary of Defense (Health Affairs).				
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)	
FIELD	GROUP	SUB-GROUP	Fort Bragg Evaluation Project, Statistical Power Analysis	
19. ABSTRACT (Continue on reverse if necessary and identify by block number) This report presents the statistical review regarding an extension of the Fort Bragg Evaluation Project by Vanderbilt University Center for Mental Health Policy. It contains an assessment of two data collection plans using power analysis. The Monte Carlo power analysis performed by Vanderbilt University is also evaluated. Based on the current short-term data collection plan submitted by the State of North Carolina, the statistical power is computed to be 80.258%. This level of power is considered high and should be adequate to meet the published Fort Bragg Evaluation Project statement of work.				
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT <input type="checkbox"/> UNCLASSIFIED/UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Scott A. Optenberg			22b. TELEPHONE (Include Area Code) (210) 221-0278	22c. OFFICE SYMBOL HSHN-A

ASSESSMENT OF TWO DATA COLLECTION
APPROACHES FOR FORT BRAGG CHILD/ADOLESCENT
MENTAL HEALTH DEMONSTRATION PROJECT
USING POWER ANALYSIS

A REPORT TO
THE ASSISTANT SECRETARY OF DEFENSE
(HEALTH AFFAIRS)

Barbara E. Wojcik, Ph.D.
Senior Scientific Statistician

PRC, Inc.
San Antonio, Texas

Catherine R. Stein, MS, GS-11
Statistician

Dr. Scott A. Optenberg, GM-15
Chief, Health Care Analysis Division

Directorate of
Health Care Studies
and
Clinical Investigation

CR 93-002
Part I - Final Report (Revised)
June 1993

UNITED STATES ARMY
MEDICAL DEPARTMENT CENTER AND SCHOOL
FORT SAM HOUSTON, TEXAS 78234-6100

TABLE OF CONTENTS

SECTION	PAGE
DISCLAIMER	i
REPORT DOCUMENTATION PAGE	ii
TABLE OF CONTENTS	iv
BACKGROUND	1
POWER ANALYSIS COMPARISON OF TWO DATA COLLECTION PLANS	1
Power Analysis Assumptions	1
Power Analysis of Short and Long-Term Plans	3
Computational Procedure for the Exact Power of the Short and Long-Term Plans	4
Additional Power Computations	6
Assessment of the Simulation Method	7
CONCLUSION	9
REFERENCES	10
DISTRIBUTION LIST	12
APPENDIX A: LETTER DATED FEBRUARY 15, 1993, FROM DR. LENORE BEHAR TO MR. LEO SLEIGHT . . .	A-1 TO A-6
APPENDIX B: "STATISTICAL POWER IN CHILD PSYCHO- THERAPY OUTCOME RESEARCH," PAPER BY C. LAMPMAN, J. DURLAK, AND A. WELLS (PRESENTED AT 1992 AMERICAN PSYCHOLOGICAL ASSOCIATION CONVENTION)	B-1 TO B-2
APPENDIX C: POWER ANALYSIS DISCUSSION AND DOCUMENTATION FROM MATERIAL SUBMITTED BY VANDERBILT UNIVERSITY, APRIL 30, 1993	C-1 TO C-7

BACKGROUND

In response to inquiries from Congressional representatives, the Acting Assistant Secretary of Defense (Health Affairs) requested that the Army document a Department of Defense (DoD) position regarding an extension of the Fort Bragg Mental Health Demonstration Project. It was requested that the Army establish a panel of Army/DoD experts (psychiatrists, psychologists, other clinicians, and clinical statisticians) to review the evaluation and other related data concerning the Demonstration Project in order to: (1) support a DoD position on the level of confidence necessary to confirm treatment results/conclusions, and (2) indicate the impact of an Army approved evaluation due date on that level of confidence.

This technical report presents an independent statistical analysis/review. No actual data from the Fort Bragg Child/Adolescent Mental Health Demonstration Project or the Fort Bragg Evaluation Project were made available. However, information contained in a letter (shown as Appendix A) written by Dr. Lenore Behar, Ph.D., Head of the Child and Family Services Branch, North Carolina Department of Human Resources, to Mr. Leo Sleight, Central Contracting Office, Department of the Army, Headquarters U.S. Army Health Services Command, Fort Sam Houston, Texas, dated February 15, 1993, was provided by Vanderbilt University. In the letter, Dr. Behar presented two data collection plans. These plans, one Short-Term and one Long-Term, differ in the number of cases collected at 'Wave 3'. The effectiveness of each plan was described by means of a power value of a statistical test for detecting differences in improvement in mental health outcomes between Demonstration and Comparison cases. In addition, a reprint of a paper submitted to the 1992 American Psychological Association Convention addressing power analysis in psychotherapy research was furnished. This paper is included as Appendix B.' Also submitted was documentation supporting the power values in Appendix A in materials attached to a letter dated April 30, 1993, written by Dr. Leonard Bickman, Ph.D., Director of the Center for Mental Health Policy, Institute for Public Policy Studies, Vanderbilt University, to LTC Thomas E. Leonard, Headquarters U.S. Army Health Services Command, Fort Sam Houston, Texas. Pertinent portions of this documentation are included as Appendix C.

POWER ANALYSIS COMPARISON OF TWO DATA COLLECTION PLANS

Power Analysis Assumptions.

In the statistical assumptions presented in Appendix A, the type of variable(s) used to measure 'improvement' between an average Demonstration case and an average Comparison case was

not defined. The paper shown in Appendix B was referenced instead, presenting the results of a meta-analysis for 12 categories of outcome measures, six each for behavioral and nonbehavioral treatments. It appears that the Fort Bragg Evaluation Project used the Appendix B paper to obtain the value of the effect size (ES) for Normed Rating Scales--Nonbehavioral Treatment outcome measures--as this value is included in Appendix A. In Appendix A (p. A-6), it is stated that the Short-Term Plan has 50% power and the Long-Term Plan of data collection would have 80% power. These levels of power were based on a simulation model submitted by Vanderbilt University (Appendix C).

The effect size (ES) index identified as d by Cohen (1988),² is the standardized difference between two population means. This equation is as follows:

$$d = \frac{m_A - m_B}{\sigma}$$

where d = ES index for t test of means,
 m_A, m_B = population means,
 and σ = standard deviation of either population
 (equal variance is assumed).

The effect size value (ES = 0.25) derived in Appendix B (p. B-2) and cited in Appendix A (p. A-5) should be used with caution for several reasons. First, this value was computed for a series of 12 sub-group samples. The Normed Rating Scale used to derive the power in Appendix A was based on a mean sample of only 33 cases. The authors of the Appendix B paper stated this problem of variability as follows (p. B-2): "The large discrepancies between sample sizes actually used and those necessary to attain an acceptable level of power in the studies shown in Table 1 make it difficult to assess how closely the obtained treatment effect sizes represent true population effects. This, in turn underscores the need for researchers to attend to power considerations when planning therapy outcome studies." When a meta-analysis is based on such a small size the probability of error is high. As a result, the mean effect size (ES = 0.25) used in Appendix A may or may not express score distances (in units of variability) for the actual variables measuring health outcome in the Fort Bragg Evaluation Project.

Secondly, there is always a risk that meta-analysis may have employed inappropriate assumptions with regard to the validity of pooling and generality. For instance, the meta-analysis may contain some bias as to how the outcome should be produced, excluding some relevant trials from analysis. In other instances, meta-analysis may use multiple results from the same study, and because the results are not independent they may

bias or invalidate the meta-analysis. In other cases, the independent studies may include different measuring techniques and definitions of variables, so the outcomes may not be comparable. In general, effect sizes in unique areas are likely to be small ($ES = 0.20$ or $ES = 0.30$), but only a pilot test would give an answer as to the probable magnitude of the ES index for the particular variable of interest in a particular situation.

The power and sample size tables (Cohen, 1988)³ for the above specified $ES = 0.25$ in Appendix A are designed to analyze the difference between means of two independent samples of the same size drawn from normal populations with equal variances (using the t test for means). If these assumptions cannot be made, which is often the case, the additional adjustments that follow are explicitly supported by Cohen (1988)⁴ and others. Computations should be performed to obtain the harmonic mean if samples of different sizes but equal variance are present, and the root mean square should be computed if two samples of the same size having unequal variances are present. If both sample sizes and variances differ, the values for power formulas from the tables cited in Appendix A may not be valid.

Since no actual data were available from the Fort Bragg Evaluation Project, this review will utilize the data used by Vanderbilt University for this analysis. Appendix A contains a comparison of the two data collection plans using power analysis. The Appendix A power analysis comparison presents the number of cases after attrition for both the Short-Term and Long-Term Plans (p. A-6). For the Short-Term Plan, 299 Demonstration cases and 150 Comparison cases were expected. The following power analysis is based on Cohen's formulas and uses the information supplied in Appendix A. This analysis is followed by a discussion of the simulation submitted by Vanderbilt University and included as Appendix C.

Power Analysis of Short and Long-Term Plans.

Under the assumption that the variances in the Demonstration and Comparison sites are equal, the harmonic mean (n) of the Demonstration sample size (n_D) and the Comparison sample size (n_C) is given by the formula (Cohen, 1988):⁵

$$n = \frac{2n_D n_C}{n_D + n_C} = \frac{2(299)(150)}{299 + 150} = \frac{89,700}{449} \approx 200.$$

The value for power of the t test of the Demonstration case mean (m_D) and the Comparison case mean (m_C) testing the null hypothesis that $m_D = m_C$ at $\alpha_1 = 0.05$ (one-tailed test) (Table 2.3.2 from Cohen, 1988)⁶ gives the following results:

for $n = 200$ and $ES = 0.20$, power = 0.64, and
 for $n = 200$ and $ES = 0.30$, power = 0.91.

The effect size, proposed in Appendix A and derived from a meta-analysis performed in Appendix B, is 0.25. A linear interpolation was performed to derive the power of the t test for $ES = 0.25$. This computation yielded a power of 0.78 for $ES = 0.25$, $\alpha_1 = 0.05$ and $n = 200$. This power of 0.78 (78%), as computed for the Short-Term Plan, is much higher than the 0.50 (50%) quoted in Appendix A. A full precision computation of the power for the Short and Long-Term Plans is presented in the next section of this report.

The Long-Term Plan projects 426 Demonstration cases and 361 Comparison cases. This harmonic mean, computed under the assumption that the variances are the same, is as follows (Cohen, 1988):⁷

$$n = \frac{2n_D n_C}{n_D + n_C} = \frac{2(426)(361)}{426 + 361} = \frac{307,572}{787} = 390.8 \approx 391.$$

Employing Table 2.3.2 in Cohen (1988),⁸ $n = 350$ yields power = 84% for $ES = 0.20$ and power = 99% for $ES = 0.30$. For $n = 400$, power = 88% for $ES = 0.20$ and power is greater than 99% for $ES = 0.30$. The linear approximation yields a power of 93.3% for $ES = 0.25$ (for $n = 391$).

Computational Procedure for the Exact Power of the Short and Long-Term Plans.

The linear interpolation to compute power, discussed on pages 3 and 4, was justified by its simplicity and by the relatively accurate values obtained. The full precision in computing the power for the Short and Long Term Plans was based on the expression (Cohen, 1988):⁹

$$z_{1-\beta} = \frac{d(n-1)\sqrt{2n}}{2(n-1) + 1.21(Z_{1-\alpha_1} - 1.06)} - Z_{1-\alpha_1}$$

where $z_{1-\beta}$ = the percentile of the standard normal distribution giving the power value
 $z_{1-\alpha_1}$ = the percentile of the standard normal distribution for α_1 significance level
 d = the effect size ES
 and n = the harmonic mean.

For the Short-Term Plan, the following information was available:

$$\begin{aligned} n &= 200 \\ \alpha_1 &= 0.05 \\ d &= 0.25 \\ z_{1-\alpha_1} &= 1.645. \end{aligned}$$

The $z_{1-\beta}$ percentile was computed under these assumptions from the above formula:

$$\begin{aligned} z_{1-\beta} &= \frac{(0.25)(200-1)\sqrt{2(200)}}{2(200-1) + 1.21(1.645-1.06)} - 1.645 \\ &= \frac{(0.25)(199)(20)}{398 + (1.21)(0.585)} - 1.645 = \frac{995}{398.708} - 1.645 \\ &= 2.496 - 1.645 = 0.851. \end{aligned}$$

The probability for this $z_{1-\beta}$ percentile was found from the Normal Curve Areas Table C (Daniel, 1988).¹⁰ This probability presents the power of the test and is equal to 80.258%. The Short-Term Plan gives a statistical power (computed with full precision) exceeding 80%.

A similar computation was performed for the Long-Term Plan under the following assumptions:

$$\begin{aligned} n &= 391 \\ \alpha_1 &= 0.05 \\ d &= 0.25 \\ z_{1-\alpha_1} &= 1.645. \end{aligned}$$

The $z_{1-\beta}$ percentile found from the same formula (Cohen, 1988)¹¹ was computed as follows:

$$\begin{aligned} z_{1-\beta} &= \frac{(0.25)(391-1)\sqrt{2(391)}}{2(391-1) + 1.21(1.645-1.06)} - 1.645 \\ &= \frac{(97.5)(27.964)}{780 + 0.70785} - 1.645 = \frac{2,726.516}{780.708} - 1.645 \\ &= 3.492 - 1.645 = 1.847. \end{aligned}$$

The power for this value of $z_{1-\beta}$ found from the Normal Curve Areas Table C (Daniel, 1988)¹² is equal to 96.78%.

Additional Power Computations.

The power analysis shown above projects that the number of cases in the Short-Term Plan is currently sufficient to draw statistically significant conclusions with high statistical power (80.258%). An additional reason for this conclusion is found by using the sample size tables provided by Cohen (1988)¹³ and deriving the sample size necessary to achieve full 80% power. Sample size tables provide data for two homogeneous normally distributed populations from which random samples of the same size were derived. The ES specified in Appendix A is 0.25. This ES level is not tabulated by Cohen (1988).¹⁴ Therefore, to find the sample size for an untabulated effect size, the following formula is used (Cohen, 1988):¹⁵

$$n = \frac{n_{.10}}{100d^2} + 1$$

where $n_{.10}$ is the sample size for desired power,
given α and $ES = 0.10$,
and d is the effect size.

In addition, if the sample sizes are not equal, one sample size is treated as if fixed, while the other is computed. When the choice is arbitrary, it is generally supported that n_c be fixed and n_D be computed. To find n_D , the following formula is used (Cohen, 1988):¹⁶

$$n_D = \frac{n_c n}{2n_c - n}$$

where n_c = fixed sample size (Comparison sites),
 n = value read from the Table 2.4.1 (Cohen,
1988)¹⁷ or computed from the previous equation,
and n_D = sample size for the Demonstration site.

With the objective to determine the Demonstration case sample size required to yield a power = 80% with $\alpha_1 = 0.05$ and $ES = 0.25$, and fixing the Comparison cases at $n = 150$ (the current level), the formula for computing n is:

$$n = \frac{n_{.10}}{100d^2} + 1 = \frac{1,237^*}{100(0.25)^2} + 1 = \frac{1,237}{6.25} + 1 \approx 198 + 1 = 199.$$

*Source: Table 2.4.1 (Cohen, 1988).¹⁸

Next, this value is put into the formula for n_D :

$$n_D = \frac{n_c n}{2n_c - n} = \frac{(150)(199)}{2(150) - 199} = \frac{29,850}{300 - 199}$$

$$= \frac{29,850}{101} = 295.54 \approx 296.$$

Consequently, 296 Demonstration site patients are needed to assure an 80% power for the test investigating the difference in mental health outcomes between Demonstration and Comparison patients (299 were projected in Appendix A).

The identical procedure was applied to the Long-Term Plan. Given that the Comparison sites consist of 361 cases, and assuming the same conditions ($\alpha_1 = 0.05$, $ES = 0.25$, power = 0.80), a sample size of 138 cases for the Demonstration site was obtained:

$$n = \frac{n_{.10}}{100d^2} + 1 = 199$$

$$n_d = \frac{n_c n}{2n_c - n} = \frac{(361)(199)}{2(361) - 199} = \frac{71,839}{722 - 199} = \frac{71,839}{523}$$

$$= 137.36 \approx 138.$$

As proposed, in Appendix A, the Long-Term Plan is projected to produce 426 Demonstration cases. Using Vanderbilt University's information taken from Appendix A, the above analysis computes only 138 cases are statistically necessary to achieve 80% power.

Assessment of the Simulation Method.

Vanderbilt University's use of the Monte Carlo simulation method to perform a power analysis in the present situation is an inappropriate application of this type of simulation. Using simulation to compute the power analysis without any information about the actual data is not an appropriate use of either simulation or power analysis. Concerning simulation, Miller and Starr (1969)¹⁹ state:

"...Simulation is not a substitute for knowledge [emphasis by authors]. This cannot be over-emphasized. Simulation is not a method, which, somehow, compensates for lack of knowledge."

In general, simulation should be treated as a technique of "last resort" (Naylor, 1971),²⁰ to be used only when analytical techniques are not available for obtaining solutions to a given model. Power analysis gives the correct probability of getting a significant result of Comparison and Demonstration site means only when the effect size is computed precisely (i.e., based on actual data from actual variables in the experiment under consideration).

The use of simulation requires complete information about the process or object. In order to simulate reasonably, the probability distributions of the variables of interest should be known. If these distributions are not known, it is impossible to simulate the process. This position is strongly emphasized by many authorities in operations research (Naylor; Ignizio and Gupta; Buffa; Smith; Banks and Carson; Gibra; and Miller and Starr).²¹ It is critical that estimates of parameters of the simulation model be derived on the basis of observations taken from the actual data. Naylor (1971)²² states:

"... There is very little to be gained by using an inadequate model to carry out simulation experiments on a computer because we would merely be simulating our own ignorance."

Since the Monte Carlo technique presented in Appendix C does not involve actual data, the results obtained from this method may be entirely misleading and not accurate. The simulation shown in Appendix C is based on assumptions regarding the effect size ($ES = 0.25$). This value, derived from meta-analysis (Appendix B, p. B-2), may not apply to real differences between the mean values of mental health outcomes for the Demonstration and Comparison sites. Another assumption (Appendix A, p. A-5), regarding the average child improvement by 0.3 SD, due to treatment and time, is only theoretical because it is not based on actual data.

As stated above, Monte Carlo simulation should only be utilized when direct data analysis cannot be performed (Gibra, 1973),²³ which is not the case with the Fort Bragg Evaluation Project. In addition, the real probability distributions of all the random variables under consideration must be given (Gibra, 1973),²⁴ a fact ignored in Appendix C. The Monte Carlo method gives only approximations to sampling distributions (Snedecor and Cochran, 1980).²⁵ To this extent, the technique itself is subject to sampling error.

Another observation about the Appendix C discussion was that the Monte Carlo method was performed only for one variable (CBCL); no other variables were used. The analysis might had different results if the other variables were considered. Finally, before any simulation model can be accepted it must be verified and validated to identify model biases and erroneous assumptions, if any. The authors of the modeling as reported in Appendix C included no such validation.

Without the use of actual data, the effect size value (derived from the meta-analysis cited in Appendix B) was used to calculate the power in this report. This effect size was recommended by the staff of the Fort Bragg Evaluation Project. Although not considered actual data, the effect size allowed for no additional bias to be created by the Monte Carlo method. The equations used to compute the power of the test of means in this report are supported by numerous authorities in power analysis (Cohen, 1988).²⁶

CONCLUSION

The power values for the directional tests computed in this study and the values given in the proposal in Appendix A are significantly different. Utilizing information available in Appendix A and a methodology well supported in the statistical literature, this study demonstrates that the Short-Term Plan would yield power exceeding 80% (80.258%) at full precision, instead of 50% as presented in Appendix A. Even using linear interpolation, a power of 78% was derived. This study demonstrates that it is unnecessary to extend the duration of the project based on power requirements; the Short-Term Plan should produce high power to demonstrate significance if the alternative hypothesis is true. The Demonstration sample size n_D needed to achieve 80% power for the Short-Term Plan ($\alpha = 0.05$, $n_C = 150$, $ES = 0.25$) equals 296 cases.

Secondly, because the standardized effect size is a computed variable, it can be modified. This modification can be achieved by any of several methods currently available to the Fort Bragg Evaluation Project staff without any project extension. Variance can be reduced, thereby allowing a decrease in sample size necessary to detect a particular level of effect size at a specified power by increasing quality control in data collection and preparation for analysis. For example, each outcome should be used in as sensitive a form as can be reliably measured (variable of interest should always be measured on a continuum, not dichotomized). Unnecessary dichotomization causes a loss of power in all analyses. Consequently, a much larger sample is necessary to achieve the same power.

Finally, as stated above, a more accurate estimate of the Fort Bragg Evaluation Project effect size is achieved when actual data is utilized and a full post hoc power analysis is conducted. The advisability of performing post hoc power analysis is strongly supported by Cohen (1988),²⁷ Rossi (1990),²⁸ Bailar (1992),²⁹ and numerous authorities on power analysis in the behavioral/medical sciences.

REFERENCES

1. Claudia Lampman, Joseph Durlak, and Anne Wells, "Statistical Power in Child Psychotherapy Outcome Research," Paper presented at the annual convention of the American Psychology Association, 1992.
2. Jacob Cohen, Statistical Power for the Behavioral Sciences (Hillsdale, NJ: Lawrence Erlbaum Associates, 1988), 20.
3. Ibid.
4. Ibid., 42.
5. Ibid., 42.
6. Ibid., 31.
7. Ibid., 42.
8. Ibid., 31.
9. Ibid., 544.
10. Wayne W. Daniel, Essentials of Business Statistics, 2nd Ed. (Boston, MA: Houghton Mifflin Co., 1988), A26-A27.
11. Cohen, 544.
12. Daniel, A26-A27.
13. Cohen, 54.
14. Ibid., 54.
15. Ibid., 53.
16. Ibid., 59.
17. Ibid., 54.
18. Cohen, 54.
19. David W. Miller and Martin K. Starr, Executive Decisions and Operations Research, 2nd Ed. (Englewood Cliffs, NJ: Prentice-Hall, 1969), 556.
20. Thomas H. Naylor, Computer Simulation Experiments with Models of Economic Systems (New York: John Wiley & Sons, 1971).
21. Thomas H. Naylor, Computer Simulation Experiments with Models of Economic Systems (New York: John Wiley & Sons, 1971); James P. Ignizio and Jatinder N. D. Gupta, Operations Research in Decision Making, with the collaboration of Gerald R. McNichols

(New York: Crane, Russak & Co., 1975); Elwood S. Buffa, Operations Management: Problems and Models, 3rd Ed. (New York: John Wiley & Sons, 1972); V. Kerry Smith, Monte Carlo Methods: Their Role for Econometrics (Lexington, MA: Lexington Books, D.C. Heath and Co., 1973); Jenny Banks and John S. Carson, II, Discrete-Event System Simulation (New York: Prentice-Hall, 1984); Isaac Gibra, Probability and Statistical Inference for Scientists and Engineers (Englewood Cliffs, NJ: Prentice-Hall, 1973); and David W. Miller and Martin K. Starr, Executive Decisions and Operations Research, 2nd Ed. (Englewood Cliffs, NJ: Prentice-Hall, 1969).

22. Naylor, 14.

23. Isaac N. Gibra, Probability and Statistical Inference for Scientists and Engineers (Englewood Cliffs, NJ: Prentice-Hall, 1973), 43.

24. Ibid.

25. George W. Snedecor and William G. Cochran, Statistical Methods, 7th Ed. (Ames, IA: Iowa State University Press, 1980), 9.

26. Cohen.

27. Ibid., 14.

28. Joseph S. Rossi, "Statistical Power of Psychological Research: What Have We Gained in 20 Years?," Journal of Consulting and Clinical Psychology 58 (1992): 646-656.

29. John C. Bailar III and Frederick Mosteller, Medical Uses of Statistics, 2nd Ed. (Boston, MA: NEJM Books, 1992), 47.

DISTRIBUTION LIST

Administrator, Defense Technical Information Center, ATTN:
DTIC-OOC (Selection), Bldg 5, Cameron Station, Alexandria,
VA 22304-6145 (2)
Director, Joint Medical Library, DASG-AAFJML, Offices of
the Surgeons General, Army/Air Force, Rm 670, 5109 Leesburg
Pike, Falls Church, VA 22041-3258 (1)
Director, The Army Library, ATTN: ANR-AL-RS (Army
Studies), Rm 1A518, The Pentagon, Washington, DC 20310 (1)
Defense Logistics Studies Information Exchange, U.S. Army
Logistics Management College, Fort Lee, VA 23801-8043 (1)
Commandant, Academy Health Science, ATTN: HSHA-Z,
Fort Sam Houston, TX 78234-6100 (1)
Stimson Library, Academy of Health Sciences, Bldg 2840,
Fort Sam Houston, TX 78234-6100 (1)
Medical Library, Brooke Army Medical Center, Reid Hall,
Bldg. 1001, Fort Sam Houston, TX 78234-6200 (1)
The Assistant Secretary of Defense (Health Affairs), The
Pentagon, Washington, DC 20301-1200 (3)
Office of the Assistant Secretary of Defense (HA), Health
Services Financing (HSF), Coordinated Care Policy, Rm 1B657,
The Pentagon, Washington, DC 20301-1200 (3)
HQ HSC (HSCL-M), ATTN: COL Beunler, Fort Sam Houston, TX
78234-6000 (3)
HQ HSC (HSAA-C), ATTN: Ms Emily Mathis, Fort Sam Houston, TX
78234-6000 (3)

Appendix F.
Health Services Command's May 21, 1993
Request for Data for Independent Power Analysis

TEL:

May 21.93 14:13 No.008 P.02



REPLY TO
ATTENTION OF

DEPARTMENT OF THE ARMY
HEADQUARTERS, UNITED STATES ARMY HEALTH SERVICES COMMAND
FORT SAM HOUSTON, TEXAS 78234-6000



May 21, 1993

Central Contracting Office

SUBJECT: Contract DADA10-89-C-0013, Fort Bragg Mental
Health Demonstration Project, Vanderbilt Data
Request

Dr. Lanore Behar
North Carolina Department of Human Resources
Division of Mental Health, Developmental
Disabilities and Substance Abuse Services
325 North Salisbury Street
Raleigh, North Carolina 27603

Dear Dr. Behar:

We have reviewed the data submitted by Vanderbilt on April 30, 1993 and find that we are in need of additional information for post hoc analysis. Request that you have Vanderbilt provide us the information identified below by close of business May 26, 1993.

The following information on the Demonstration and Comparison populations is requested:

1. All currently collected patient enrollment and associated data to include all variables needed to identify patients, their demographic characteristics, diagnoses, etc.
2. All currently collected workload data, e.g., treatment beginning and ending dates, type of treatment provided, number of treatment sessions, data on premature withdrawals and their dates of drop-out when applicable, etc.
3. All currently collected data resulting from tests administered to evaluate and monitor patients prior to treatment, while in treatment and following treatment to include all data from all Waves (I, II, and III).
4. All summary data (summed normed scale scores) that have been computed by project staff describing mental health outcome.
5. If CHAMPUS claims data provided to Vanderbilt University through PASS has been modified for data analysis, it is requested that these modified data be provided.
6. All variable definitions, lengths and data file layouts.

TEL:

May 21, 93 14:13 No.008 P.03

-2-

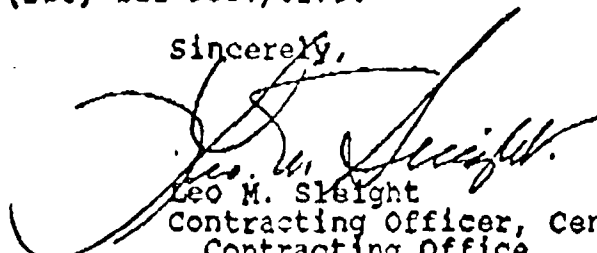
Data transfer should be made using electronic medium employing standard ASCII files. Acceptable mediums include floppy nine-track tapes or 3480 cartridge.

If you can not provide all of this data within the required time, please provide us with a written explanation detailing what data can not be provided immediately, why this can not be provided and the earliest date you can provide the requested data. You are reminded that the Army has paid for this data and has a right to receive it whenever it is requested.

Please provide us with a breakdown of any costs associated with providing this data.

If you have any further questions please contact Ms. Joyce Nadeau (210) 221-9397/0179.

Sincerely,



Leo M. Sleight
Contracting Officer, Central
Contracting Office

Copies Furnished:

Mr. James Newman, Contracting Officer's Representative,
Womack Army Medical Center, Fort Bragg, North Carolina
28307-5000

Lieutenant Colonel Leonard, Headquarters, U.S. Army
Health Services Command, Attention: HSCL-M, Fort Sam
Houston, Texas 78234

Dr. Leonard Bickman, Project Director, Vanderbilt
University, Institute for Public Policy Studies, Box
7701, Station B, Nashville, Tennessee

Appendix G.
Vanderbilt's Power Analysis of the Evaluation

VANDERBILT UNIVERSITY



NASHVILLE, TENNESSEE 37235

TELEPHONE (615) 322-7311

Institute for Public Policy Studies • Box 7701, Station B • Direct phone 322-8435

FAX 322-7049

*Center for Mental
Health Policy*

May 10, 1993

L. Colonel T. E. Leonard
Chief, Program Branch, CCD
Building 2792, Room 320
Army Health Services Command
Ft. Sam Houston TX 78234-6000

Dear Col. Leonard:

The enclosed paper, "Power analysis of the Ft. Bragg Evaluation Project: Technical details of a practical Monte Carlo power analysis" summarizes the power analysis in a form that can be reviewed by a statistician.

According to this analysis, the project needs over 800 wave 3 cases for 80% power (at the standard α level of 5%).

Sincerely,

Leonard Bickman
Director

Power Analysis of the Ft. Bragg Evaluation Project: Technical Details of a Practical Monte Carlo Power Analysis

Warren Lambert, PhD¹

The Ft. Bragg Evaluation Project required a power analysis to determine how many subjects would be required to measure differences between the Demonstration and Comparison across time. When the research is costly, assessing too many cases would waste money; too few cases would jeopardize the entire project because the actual outcome could not be distinguished from differences occurring by chance. Since the Ft. Bragg Evaluation Project's data did not fit elementary models in standard texts, a Monte Carlo model was used. Data sets were generated according to explicit assumptions and analyzed in the same way actual data might be analyzed.

Power analysis determines how many subjects are needed in a study to have the statistical power needed to detect effects (if they occur in nature). Such a determination is particularly necessary in large research projects in which data are costly to collect and analyze. If too much data is gathered, then money is wasted; if too little, the whole research project is wasted when it fails to detect the effects it was designed to measure.

Simple projects can use simple techniques available in standard texts to calculate the power of their experimental design. The most common method of power analysis is to find one's design in a power analysis text, such as Cohen (1988) or Lipsey (1990), choose the appropriate effect size, and look up the power from power curves or power tables. Computer programs, such as BMDP Solo (Hintz, 1991) automate the process and draw power curves to fit a particular situation. While this "look-up" approach works well for common problems, such as the two group t-test, advanced large-scale evaluations often have features not found in standard look-up tables.

For example, the Ft. Bragg Evaluation Project presents the following characteristics, all of which are important in determining the study's power:

1. Two experimental groups were subjected to two forms of treatment (called "Demonstration" and "Comparison");
2. Clinical status was measured by continuous variables (such as the Achenbach CBCL total score) at three waves: admission (Wave 1), and 6- and 12-month follow-up (Waves 2 and 3);

¹Center for Mental Health Policy, Vanderbilt University, 110 21st Avenue South #1100, Nashville TN 37235 (615) 3433-1895. Fax (615) 322-7049. "Lambert@UANSV3.Vanderbilt.Edu"

3. Wave 1 scores had a mean of t-score 65. All scores have a standard deviation (*SD*) of about 10. Achenbach's norms suggest that normal children not in treatment have a mean of 50 and a *SD* of 10).

4. On the average, patients in both groups improve due to [time + treatment + regression to the mean]. This effect is of little interest in this study, since we are interested in difference between treatment methods across time, but the wave effect on all subjects had to be included in a complete data model. The effect size of [time + treatment + regression to the mean] was 0.30 *SD*.

5. We hypothesized an effect size² in which all patients improved an average of 0.30 *SD* (from t-score = 65 to 62) and patients in the Demonstration improved another 0.25 *SD* (from 62 to 59.5) by Wave 3. This means that a Comparison child having a score of t-score 65 on Wave 1 intake would have a score of $65.0 - 3.0 = 62.0$ on Wave 3 one year later. A Demonstration child having a score of 65 on Wave 1 intake would have a score of $65.0 - 3.0 - 2.5 = 59.5$ on Wave 3 one year later. This effect size (2.5 points or 0.25 *SD*) was chosen, rather than a larger one, because 1) the Ft. Bragg Demonstration is a mental health system study, not a focused university-based study of a well-defined treatment vs. a well-defined nontreatment condition; and, 2) Many patients in both groups were evaluated on intake, did not return for treatment, but were evaluated in the study. These "nontreated" cases cannot be ignored when we study a mental health system, but such individuals dilute the larger effects in patients who receive regular treatment for a year or more.

6. The number of subjects will be unbalanced, because larger numbers of subjects have been recruited in the Demonstration.

7. Correlations between Wave 1, Wave 2, and Wave 3 would be $r(1,2) = 0.50$ and $r(1,3) = 0.25$. These cross-wave correlations³ occurred when a child's status at intake carries carry over somewhat to Wave 2. While the data are not yet in, the correlation between waves (1, 3) is probably less than the correlation between waves (1, 2) and between waves (2, 3). This persistence, an autoregressive effect in which scores carry over time, can make ordinary least squares statistics show significance when effects are actually due to chance. Appendix A shows cross-wave correlations for actual CBCL data taken from children (not computer-generated).

²Effect size of the effect of time and the effect of treatment across time are hypothetical. You don't know the actual effect size until the research is finished.

³Actual data suggests that the adjacent wave correlation is around 0.5 or 0.6; the reduction to 0.25 on nonadjacent waves is an educated guess.

The foregoing list of technical requirements for power analysis was too intricate to permit "look-up" solutions. Therefore, simulation was used. Scores were generated by computer programs, tested to see if they meet the requirements stated above, and then analyzed to see how many errors (false positives and false negatives) were produced by the analysis. This process is analogous to repeating the entire Ft. Bragg Evaluation Project almost 10,000 times with different numbers of subjects to see how the results actually turn out, except that the data were generated by computer according to the list of assumptions.

By analyzing scores with known characteristics, including an experimental effect, a large number of times and counting how many times the effect was detected or missed, estimates of power were observed as simple counts. For example, if in 150 analyses each with 820 subjects and the experimental effect is detected 79% of the time, one would conclude that the estimated power is 79% at $N = 820$ subjects, that the β -error rate is 21%, i.e. 21% of the time an effect exists and we fail to detect it).

Method: The Statistical Details

This technical section describes the main steps used in calculating scores in language that statistical programmers can understand. The text following explains the computer code used to generate scores; the full computer program appears as Appendix B.

The Monte Carlo power analysis studied the power of a single variable treated in a repeated measures ANOVA⁴. While a simple repeated measures interaction-F was probably the most powerful test of differences in mental status across time, Greenhouse-Geisser correction or MANOVA may be necessary because of the lack of compound symmetry of the cross-wave correlations (item 7 above). The more rigorous MANOVA measure might be less powerful, but we needed to know whether the loss in power was negligible or serious.

Stating parameters

The following SAS code states some basic parameters used to generate scores:

```
let Timefect = -0.30;  
%let Sitefect = -0.25;  
%let Rand = Rannor(0);
```

⁴Covariance analysis using Wave 1 as a covariate was considered, but the Ft. Bragg Demonstration is not a randomized experiment, and significant pretest differences exist on some clinical variables important to the outcome analysis. A MANOVA on a list of 32 outcome variables with $N = 284$ was significant ($F(32, 251) = 3.1, p < 0.001$). With Wave 1 $N = 396$, the following summary variables had significant univariate t-tests between sites: $p < .001$: CAFAS role performance, PCAS major depression, PCAS total dysthymia. $p < .01$: CAFAS behavior toward self, CBCL school, PCAS conduct disorder symptoms, PCAS overanxious, and PCAS total endorsed all items. $p < 0.05$: CAFAS substance use.

The "time effect" statement means that the effect of time and treatment was $-0.30 SD$ per patient for all patients, that is on the average they got $0.30 SD$ better (indicated by lower scores). The "site effect" statements means that Demonstration patients get an additional $0.25 SD$ better than Comparison patients by Wave 3. This $0.25 SD$ is the minimum effect size that the study is obligated to detect; if the actual effect were smaller, we are willing to report "the Demonstration's effect was negligible." The "rand" parameter means that scores will be generated with normally distributed number generators; common statistical analyses assume normally distributed data. Clinical scores from well-designed instruments, such as CBCL totals, actually are approximately normal in distribution.

```
* *****constants;
AR1      = 0.5000;
withcons = 0.1500;
wavecons = 1.0000;
varicons = 0.8100;
errconst = 0.3500;
```

The constants defined above were necessary to create scores with the proper variance and cross-wave correlation. The appropriateness of these values can be judged by the scores they produce.

Generating a score (for example a Wave 3 score)

The following equations show how the third wave score is generated from its components:

```
Xa3 = (withcons * truemith) + (wavecons * truewav4);      [1]
Xa3 = Xa3 + (AR1 * Xa2);                                   [2]
Xa3 = (varicons * (Xa3 + (errconst * err4)));              [3]
Xa3 = Xa3 + (1.0 * &Timeeffect);                           [4]
```

The first line adds true scores for the within-subject and per-Wave 3 components. The next line adds an autoregressive component (part of the Wave 2 score that carries over into Wave 3). The third line contains error variance and parameter "varicons" needed to make the overall variance come to 1.0. The fourth line adds the effect of [time + treatment + regression] for all subjects.

After the basic score is built above, the "subject" is assigned to a treatment or control group based on a random number. In this case, two-thirds of the cases are Demonstration and one-third (0.3333) are Comparison cases.

```
GroupNum = uniform(131161);
if (GroupNum < 0.3333) then site = "Comp";
                        else site = "Demo";
```

After random assignment, the effect of treatment is added, but only for cases in the Demonstration. The effect of time and of site is complete on Wave 3 and 50% complete on Wave 2 according to this model.

```

if site = "Demo" then do
    Xa2 = Xa2 + (0.5 * &Siteeffect);
    Xa3 = Xa3 + (1.0 * &Siteeffect);
end;

```

The foregoing model produces scores with a mean of zero and a standard deviation of 1 (except for the effects of site and time). The following linear transformation creates scores with a base mean t-score of 65 and a standard deviation of 10 offset by the effects of time and treatment.

```

Xa1 = 65 + (10 * Xa1); Xa2 = 65 + (10 * Xa2); Xa3 = 65 + (10 * Xa3);
Xn1 = 65 + (10 * Xn1); Xn2 = 65 + (10 * Xn2); Xn3 = 65 + (10 * Xn3);

```

The scores Xa1 Xa2 Xa3 are Wave 1-2-3 scores assuming that the null hypothesis (Ho) is false and that the alternative (Ha) is true (i.e., that there is an effect for site). The scores Xn1 Xn2 Xn3 are Wave 1-2-3 scores assuming that the null hypothesis is true and that the alternative is false (i.e. that there is no effect of site). The alternative scores let us test power: when there is an effect, what percent of the time will we in fact detect it? The null scores let us test false positives: when we set the $p(\alpha) = 0.05$ are the statistical tests performing as advertised (yielding 5% false positives), or is our imperfect autoregressive data structure causing standard tests to give false results?

After all these calculations, scores are saved in a SAS file for analysis. This file contains Wave 1, 2, and 3 scores given an effect size of 0.25 SD per subject (Ha: Xa1 Xa2 Xa3) and Wave 1, 2, and 3 scores given an effect size of zero (Ho: Xn1 Xn2 Xn3) and a value for site ("Demo" or "Comp").

When we generate a large number of these scores (e.g., 200,000 cases) descriptive statistics tell us whether the model was in fact accurate in creating data according to our assumptions.

	ALL			SITE					
				Comp			Demo		
	MEAN	STD	N	MEAN	STD	N	MEAN	STD	N
XA1 Wave 1 Ha: Site effect	65.0	9.9	200000	65.0	9.9	100036	65.0	10.0	99964
XA2 Wave 2 Ha	62.9	10.0	200000	63.5	10.0	100036	62.2	10.0	99964
XA3 Wave 3 Ha	60.8	10.2	200000	62.1	10.1	100036	59.5	10.1	99964
XN1 Wave 1 Ho: No effect	65.0	9.9	200000	65.0	9.9	100036	65.0	10.0	99964
XN2 Wave 2 Ho	63.5	10.0	200000	63.5	10.0	100036	63.5	10.0	99964
XN3 Wave 3 Ho	62.0	10.1	200000	62.1	10.1	100036	62.0	10.1	99964

This table shows that the standard deviation was about ten in all cells of the simulated study. Both groups started with a mean of t-score of 65.0. The Comparison cases went down almost 3 points⁵ (0.30 SD) due to [time + treatment + regression]. The Demonstration cases, on average, went down to a t-score of 59.5, or about 3 points for [time + treatment] and an additional 2.5 points for benefits unique to the Demonstration. $[65 - 59.5 = 5.5 = 3.0 + 2.5]$.

A correlation matrix shows the cross-wave correlations for computer-generated scores:

Pearson Correlation Coefficients				N = 200000	
	XA1	XA2	XA3		
XA1	1.00000				
XA2	0.48209	1.00000			
XA3	0.26542	0.49034	1.00000		

The correlation matrix shows a moderate $r = 0.48$ or 0.49 between adjacent waves; this correlation goes down to $r = 0.27$ on nonadjacent waves. These nonconstant cross correlations raises the possibility that the data lack the compound symmetry that ANOVA requires. Longitudinal mental health data ordinarily exhibits this pattern of diminishing carry-over, which raises the concern that ordinary least squares statistics (such as ANOVA) will erroneously report that chance differences are significant. Appendix A contains real CBCL data from children; this data illustrates the close resemblance between real data and the data generated in Monte Carlo power analysis.

The following statistics suggest that the simulation data were normally distributed.

Moments			
N	10000		
Mean	64.8521		
Std Dev	10.01939	Variance	100.3881
Skewness	0.032833	Kurtosis	-0.0008

The descriptive statistics suggest that the simulation was successful in generating scores that met the assumptions. Having generated the scores, it is straightforward to analyze them in the same way that real data might be analyzed⁶.

⁵"Almost" not exactly. Like real scores, computer generated scores are random variables and our observations are always approximate, never exact, even if we have 100,000 cases.

⁶The real analysis might be doubly multivariate (i.e. a repeated measures MANOVA on a list of variables). The present example of a single-variable MANOVA + ANOVA seemed sufficiently complex for power analysis. However, the univariate model presented here may be more powerful than a MANOVA on a list of variables.


```

%macro glm1;
    proc glm;
        classes SITE;
        model Xa1 Xa2 Xa3 = SITE/nouni;
        repeated wave 3;
    run; quit;

    proc glm;
        classes SITE;
        model Xn1 Xn2 Xn3 = SITE/nouni;
        repeated wave 3;
    run; quit;
%mend glm1;

```

These analyses describe a repeated measures ANOVA with both univariate repeated measures (ANOVA) and multivariate (MANOVA) tests of significance. The univariate ANOVA assumes symmetric covariances; the MANOVA does not. Both ANOVA and MANOVA are ordinary least squares procedures; neither is robust against large time series (autoregressive) effects. However, the damage may be small in a limited time series with only three waves. The "Ho = no effect" group was included to make sure the planned analyses reported significance levels accurately, i.e. that $\alpha = 0.05$ was not distorted by the cross-correlated data.

After data are generated and analyzed, SAS records the results in a printout file. When thousands of ANOVA's are run, this file becomes too large to read the results, but not too large for another program to "condense" the printout (i.e., delete everything but number of cases and the probability that the results were due to chance). A sample of the output follows:

```

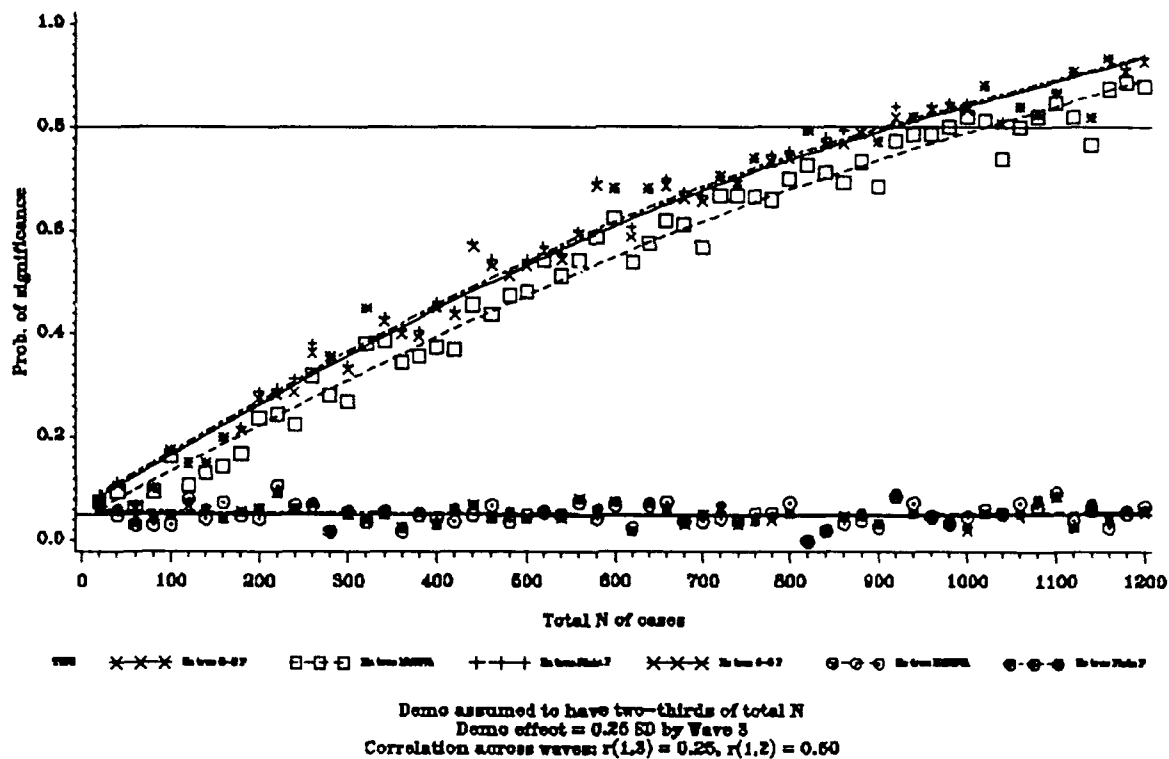
1  Number of observations in data set = 800
2  SAS
3  Dependent Variable      XA1      XA2      XA3
4  the Hypothesis of no WAVE*SITE Effect
5  Pillai's Trace          3.2232          2          797      .04035
6  Source: WAVE*SITE
7  F Value    Pr > F      G - G      H - F
8  3.90      .02038      .02199      .02185
9  Greenhouse-Geisser Epsilon = 0.9567
10 Huynh-Feldt Epsilon = 0.9601
1  Number of observations in data set = 800
2  SAS
3  Dependent Variable      XN1      XN2      XN3
4  the Hypothesis of no WAVE*SITE Effect
5  Pillai's Trace          0.0903          2          797      .91369
6  Source: WAVE*SITE
7  F Value    Pr > F      G - G      H - F
8  0.08      .91992      .91277      .91336
9  Greenhouse-Geisser Epsilon = 0.9567
10 Huynh-Feldt Epsilon = 0.9601

```

The example shows twenty data lines summarizing two of the nearly ten thousand analyses. In the first example, there were $N = 800$ cases (two-thirds in the Demonstration). There were significant effects of the Demonstration according to MANOVA ($p(\text{Pillai's}) = .04035$ and also according to ordinary F ($p = .02038$), Greenhouse-Geisser corrected F ($p = .02199$) and Huynh-Feldt corrected F ($p = .02185$). This means that the amount of change across waves differed by site. Because the variables were Xa1 Xa2 and Xa3 we know that this was an H_a case with real effects built in. In the second example of $N = 800$, the effects were nonsignificant ($p > 0.05$). Because the variables were Xn1 Xn2 and Xn3, we know that this was an H_o case with no site effect built in.

Results

Fig. 1. Statistical Power by Total N of Subjects for 9,364 outcome analyses



The power analysis was done by analyzing 9364 simulated data sets to produce Figure 1, a power curve as a function of total N . Raw data for these analyses appear in Appendices C and D. The horizontal line near the 5% level shows the outcome of 9364 data sets in which the site effect was zero (i.e. the null hypothesis, H_o , was true). Gratifyingly, we see that no matter the N , all three forms of repeated measures analysis produces very close to 5% false positives. This means that all three tests are working as they should, and that even old-

fashioned repeated measures ANOVA is not disturbed by moderate autoregression across three trials⁷.

The diagonal curve in Figure 1 is the power curve, the relative frequency of detecting results when in fact there is an additional improvement of 0.25 *SD* per patient by site. This curve reaches 80% before $N = 900$ for ANOVA and before $N = 1000$ for MANOVA. In other words, when there are almost 900 nonmissing cases (600 Demonstration and 300 Comparison), then the chance of false negatives (finding no effect when there is an effect) is 20%.

Discussion

In the 2/15/93 report to Mr. Sleight, Vanderbilt estimated that the "longer plan" would accumulate $426 + 361 = 787$ complete cases and have 80% power to detect an effect of 0.25 *SD* per case on Wave 3. The calculations presented above agree approximately with the 2/15/93 estimates, indicating that we can expect a power of 74% at $N = 787$ cases. If the present estimate, based on a much larger number of calculated cases, is the more accurate, then with $N = 787$ cases we would need a slightly stronger effect than 0.25 *SD* to have 80% power to detect a Demonstration effect of 0.25 *SD* at Wave 3.

References

- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, N.J.: Erlbaum.
- Hintz, J. (1991). *Solo Statistical System Power Analysis*. Los Angeles, CA: BMDP Statistical Software.
- Lipsey, M.W. (1990). *Design Sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage Publications.

⁷The original power analysis submitted to Mr. Sleight on 2/15/93 used only ANOVA, not MANOVA. MANOVA was added in response to technical criticism that the ANOVA might not be accurate in this situation. The MANOVA results are for reference, and we continue to use repeated measures ANOVA for the official estimate.

Appendix A:
Means and across-time correlations among real CBCL global scores
separated by six months

	SITE					
	DEMONSTRA			COMPARISON		
	MEAN	STD	N	MEAN	STD	N
T--TOTAL PROBLEM pre	66.07	9.55	245	66.02	10.64	114
T--TOTAL PROBLEM post	57.77	11.80	202	60.00	11.26	104
T--TOTAL INTERNALIZING pre	63.43	11.56	245	63.63	11.57	114
T--TOTAL INTERNALIZING post	55.43	12.25	202	56.68	11.34	104
T--TOTAL EXTERNALIZING pre	64.93	10.55	245	64.82	11.76	114
T--TOTAL EXTERNALIZING post	58.03	11.93	202	60.75	11.66	104

Pearson Coefficients⁸ / Prob > |R| under Ho: Rho=0 / Number of Observations

	X1CBCL29	X1CBCL31	X1CBCL33
X2CBCL29	0.55032		
T--TOTAL PROBLEM	4.E-23		
	275		
X2CBCL31		0.53246	
T--TOTAL INTERNALIZING		2.E-21	
		275	
X2CBCL33			0.61447
T--TOTAL EXTERNALIZING			6.E-30
			275

⁸Wave 3 data unavailable May 5, 1993.

Appendix B.

Complete Program used to generate and analyze scores in the simulation

```

options number;

%let ProbName = "effect size = 0.25 SD";
%let Siteeffect = -0.25;
%let Timeeffect = -0.30;
%let Rand = Rannor(0);

%Macro MakeData;
*x "c:\util\0 d:";
*x "echo N=&Total_N";

Data MonteCar(keep = Xa1 Xa2 Xa3 Xn1 Xn2 Xn3 site);
*****constant*****
AR1 = 0.5000; /* .33 0.50 */
withcons = 0.1500; /* .33 0.50 */
wavecons = 1.0000; /* .67 1.50 */
varicons = 0.8100; /* 1.10 0.50 */
errconst = 0.3500; /* .25 0.50 */

do i = 1 to &Total_N by 1;
  truweth = &Rand; truwav0 = &Rand; truwav1 = &Rand;
  truwav2 = &Rand; truwav3 = &Rand; truwav4 = &Rand;

  enr1 = &Rand; enr2 = &Rand; enr3 = &Rand; enr4 = &Rand;

  X00 = (withcons * truweth) + (wavecons * truwav0);
  X01 = (withcons * truweth) + (wavecons * truwav1);
  Xa1 = (withcons * truweth) + (wavecons * truwav2);
  Xa2 = (withcons * truweth) + (wavecons * truwav3);
  Xa3 = (withcons * truweth) + (wavecons * truwav4);

  X01 = X01 + (AR1 * X00);
  Xa1 = Xa1 + (AR1 * X01);
  Xa2 = Xa2 + (AR1 * Xa1);
  Xa3 = Xa3 + (AR1 * Xa2);

  Xa1 = (varicons * (Xa1 + (errconst * enr2)));
  Xa2 = (varicons * (Xa2 + (errconst * enr3)));
  Xa3 = (varicons * (Xa3 + (errconst * enr4)));

  Xa2 = Xa2 + (0.5 * &Timeeffect);
  Xa3 = Xa3 + (1.0 * &Timeeffect);

  Xn1 = Xa1; Xn2 = Xa2; Xn3 = Xa3;

  GroupNum = uniform(131161);
  if (GroupNum < &Split) then do
    site = "Comp";
  end;
  else do
    site = "Demo";
  end;

  if site = "Demo" then do
    Xa2 = Xa2 + (0.5 * &Siteeffect);
    Xa3 = Xa3 + (1.0 * &Siteeffect);
  end;

  Xa1 = 65 + (10 * Xa1); Xa2 = 65 + (10 * Xa2); Xa3 = 65 + (10 * Xa3);
  Xn1 = 65 + (10 * Xn1); Xn2 = 65 + (10 * Xn2); Xn3 = 65 + (10 * Xn3);

  length Xa1 Xa2 Xa3 Xn1 Xn2 Xn3 4;
  output;
end;

run;

%Mend MakeData;

```

```
%macro describe;
options linesize = 120;
proc tabulate f = 5.2 data = MonteCar;
  class SITE;
  var Xa1 Xa2 Xa3 Xn1 Xn2 Xn3;
  table (Xa1 Xa2 Xa3 Xn1 Xn2 Xn3),
    (all SITE)*(mean*f=8.4 std*f=8.4 N*f=6.0);
  title &ProbName;
run;

proc tabulate f = 5.2 data = MonteCar;
  class SITE;
  var Xa1 Xa2 Xa3 Xn1 Xn2 Xn3;
  table (Xa1 Xa2 Xa3 Xn1 Xn2 Xn3),
    (all SITE)*(mean*f=5.1 std*f=5.1 N*f=6.0);
  title &ProbName;
run;

proc corr nosimple;
  var Xa1 Xa2 Xa3;
run;
proc corr nosimple;
  var Xn1 Xn2 Xn3;
run;

proc univariate data = MonteCar;
  var Xa1 Xa2 Xa3 Xn1 Xn2 Xn3;
run;

%mend describe;

%macro glm1;
options linesize = 80;
%MakeData;
proc glm;
  classes SITE;
  model Xa1 Xa2 Xa3 = SITE/nouni;
  repeated wave 3;
run; quit;

proc glm;
  classes SITE;
  model Xn1 Xn2 Xn3 = SITE/nouni;
  repeated wave 3;
run; quit;

%mend glm1;

%macro glm10;
%glm1; %glm1; %glm1; %glm1; %glm1; %glm1; %glm1; %glm1; %glm1;
%mend glm10;

%macro glm50;
%glm10; %glm10; %glm10; %glm10; %glm10;
%mend glm50;

* *****
* Set numbers prior to running a problem
* *****;
%macro Run10all;

%let Split = 0.333333 ;

%let Total_N = 20; %glm10;
%let Total_N = 40; %glm10;
%let Total_N = 60; %glm10;
%let Total_N = 80; %glm10;

%let Total_N = 100; %glm10;
%let Total_N = 120; %glm10;
%let Total_N = 140; %glm10;
%let Total_N = 160; %glm10;
%let Total_N = 180; %glm10;

%let Total_N = 200; %glm10;
%let Total_N = 220; %glm10;
%let Total_N = 240; %glm10;
%let Total_N = 260; %glm10;
%let Total_N = 280; %glm10;

%let Total_N = 300; %glm10;
%let Total_N = 320; %glm10;
%let Total_N = 340; %glm10;
%let Total_N = 360; %glm10;
%let Total_N = 380; %glm10;
```

%let Total_N = 400; %glm10;
%let Total_N = 420; %glm10;
%let Total_N = 440; %glm10;
%let Total_N = 460; %glm10;
%let Total_N = 480; %glm10;

%let Total_N = 500; %glm10;
%let Total_N = 520; %glm10;
%let Total_N = 540; %glm10;
%let Total_N = 560; %glm10;
%let Total_N = 580; %glm10;

%let Total_N = 600; %glm10;
%let Total_N = 620; %glm10;
%let Total_N = 640; %glm10;
%let Total_N = 660; %glm10;
%let Total_N = 680; %glm10;

%let Total_N = 700; %glm10;
%let Total_N = 720; %glm10;
%let Total_N = 740; %glm10;
%let Total_N = 760; %glm10;
%let Total_N = 780; %glm10;

%let Total_N = 800; %glm10;
%let Total_N = 820; %glm10;
%let Total_N = 840; %glm10;
%let Total_N = 860; %glm10;
%let Total_N = 880; %glm10;

%let Total_N = 900; %glm10;
%let Total_N = 920; %glm10;
%let Total_N = 940; %glm10;
%let Total_N = 960; %glm10;
%let Total_N = 980; %glm10;

%let Total_N = 1000; %glm10;
%let Total_N = 1020; %glm10;
%let Total_N = 1040; %glm10;
%let Total_N = 1060; %glm10;
%let Total_N = 1080; %glm10;

%let Total_N = 1100; %glm10;
%let Total_N = 1120; %glm10;
%let Total_N = 1140; %glm10;
%let Total_N = 1160; %glm10;
%let Total_N = 1180; %glm10;

%let Total_N = 1200; %glm10;

%mend Run10all;

%Macro Run100;
%Run10all; %Run10all;%Run10all;%Run10all;%Run10all;
%Run10all; %Run10all;%Run10all;%Run10all;%Run10all;
%Mend Run100;

%Run100; %Run100;%Run100; %Run100;

endsas;

%let Split = 0.50 ;
%let Total_N = 200000;
%MakeData; %Describe;

endsas;

Appendix C.

Mean Power for 9364 analyses Estimates: Effect size = 0.25 SD

OBS	EFFECTYP	TOTAL N	FREQ	MEANPIL	MEANF	MEANGG
1	Ha true	20	160	0.0750	0.0875	0.0688
2	Ha true	40	160	0.0937	0.1125	0.1062
3	Ha true	60	160	0.0625	0.0688	0.0625
4	Ha true	80	160	0.0937	0.1000	0.1000
5	Ha true	100	160	0.1625	0.1750	0.1750
6	Ha true	120	160	0.1062	0.1500	0.1500
7	Ha true	140	160	0.1313	0.1500	0.1500
8	Ha true	160	160	0.1437	0.2000	0.2000
9	Ha true	180	160	0.1688	0.2187	0.2125
10	Ha true	200	160	0.2375	0.2875	0.2750
11	Ha true	220	160	0.2437	0.2938	0.2812
12	Ha true	240	160	0.2250	0.3125	0.2875
13	Ha true	260	160	0.3187	0.3812	0.3625
14	Ha true	280	160	0.2812	0.3563	0.3563
15	Ha true	300	160	0.2687	0.3375	0.3312
16	Ha true	320	160	0.3812	0.4500	0.4500
17	Ha true	340	160	0.3875	0.4313	0.4250
18	Ha true	360	160	0.3437	0.4062	0.4000
19	Ha true	380	160	0.3563	0.4062	0.3937
20	Ha true	400	160	0.3750	0.4625	0.4500
21	Ha true	420	160	0.3688	0.4437	0.4375
22	Ha true	440	160	0.4562	0.5750	0.5687
23	Ha true	460	160	0.4375	0.5437	0.5312
24	Ha true	480	160	0.4750	0.5188	0.5125
25	Ha true	500	160	0.4813	0.5437	0.5312
26	Ha true	520	160	0.5437	0.5687	0.5625
27	Ha true	540	160	0.5125	0.5563	0.5437
28	Ha true	560	160	0.5437	0.6000	0.5937
29	Ha true	580	160	0.5875	0.6937	0.6875
30	Ha true	600	160	0.6250	0.6813	0.6813
31	Ha true	620	160	0.5375	0.6062	0.5875
32	Ha true	640	160	0.5750	0.6813	0.6813
33	Ha true	660	160	0.6188	0.7000	0.6875
34	Ha true	680	160	0.6125	0.6750	0.6625
35	Ha true	700	160	0.5687	0.6687	0.6562
36	Ha true	720	160	0.6687	0.7063	0.7063
37	Ha true	740	154	0.6688	0.6983	0.6983
38	Ha true	760	150	0.6667	0.7400	0.7400
39	Ha true	780	150	0.6600	0.7467	0.7333
40	Ha true	800	150	0.7000	0.7533	0.7400
41	Ha true	820	150	0.7267	0.7933	0.7933
42	Ha true	840	150	0.7133	0.7800	0.7667
43	Ha true	860	150	0.6933	0.7933	0.7667
44	Ha true	880	150	0.7333	0.7867	0.7867
45	Ha true	900	150	0.6867	0.7733	0.7733
46	Ha true	920	150	0.7733	0.8400	0.8200
47	Ha true	940	150	0.7867	0.8200	0.8200
48	Ha true	960	150	0.7867	0.8400	0.8333
49	Ha true	980	150	0.8000	0.8467	0.8400
50	Ha true	1000	150	0.8200	0.8467	0.8333
51	Ha true	1020	150	0.8133	0.8800	0.8800
52	Ha true	1040	150	0.7400	0.8133	0.8067
53	Ha true	1060	150	0.8000	0.8400	0.8400
54	Ha true	1080	150	0.8200	0.8267	0.8267
55	Ha true	1100	150	0.8467	0.8667	0.8667
56	Ha true	1120	150	0.8200	0.9067	0.9067
57	Ha true	1140	150	0.7667	0.8200	0.8200
58	Ha true	1160	150	0.8733	0.9333	0.9333
59	Ha true	1180	150	0.8867	0.9133	0.9067
60	Ha true	1200	150	0.8800	0.9333	0.9267

Notes:

Ha true means that there really was an effect of Demonstration treatment.

Total N refers to the N of subjects (children) in the calculated model.

Freq refers to the number of times the model was created and analyzed. The freqs are uneven because the computer ran all night and was stopped the next morning.

"Meanpil" stands for the mean power by Pillai's trace MANOVA. Mean F and MeanGG stand for classical formula repeated measures F and Greenhouse-Geisser F corrected for possibly asymmetric covariances.

Appendix D.
Mean Power Estimates for 9364 analyses: Effect size = 0.00 *SD*
(Null hypothesis is true)

61	Ho true	20	160	0.0750	0.0750	0.0625
62	Ho true	40	160	0.0500	0.0625	0.0563
63	Ho true	60	160	0.0313	0.0375	0.0313
64	Ho true	80	160	0.0313	0.0437	0.0437
65	Ho true	100	160	0.0313	0.0500	0.0500
66	Ho true	120	160	0.0812	0.0750	0.0688
67	Ho true	140	160	0.0437	0.0625	0.0563
68	Ho true	160	160	0.0750	0.0437	0.0437
69	Ho true	180	160	0.0500	0.0563	0.0563
70	Ho true	200	160	0.0437	0.0625	0.0625
71	Ho true	220	160	0.1062	0.0937	0.0937
72	Ho true	240	160	0.0688	0.0625	0.0625
73	Ho true	260	160	0.0688	0.0750	0.0688
74	Ho true	280	160	0.0188	0.0188	0.0188
75	Ho true	300	160	0.0563	0.0563	0.0500
76	Ho true	320	160	0.0375	0.0437	0.0375
77	Ho true	340	160	0.0563	0.0563	0.0500
78	Ho true	360	160	0.0188	0.0250	0.0250
79	Ho true	380	160	0.0500	0.0563	0.0500
80	Ho true	400	160	0.0437	0.0313	0.0313
81	Ho true	420	160	0.0375	0.0625	0.0500
82	Ho true	440	160	0.0500	0.0688	0.0688
83	Ho true	460	160	0.0688	0.0437	0.0437
84	Ho true	480	160	0.0375	0.0563	0.0437
85	Ho true	500	160	0.0500	0.0437	0.0437
86	Ho true	520	160	0.0563	0.0563	0.0563
87	Ho true	540	160	0.0500	0.0500	0.0437
88	Ho true	560	160	0.0750	0.0812	0.0812
89	Ho true	580	160	0.0437	0.0625	0.0500
90	Ho true	600	160	0.0688	0.0750	0.0750
91	Ho true	620	160	0.0250	0.0188	0.0188
92	Ho true	640	160	0.0688	0.0750	0.0688
93	Ho true	660	160	0.0750	0.0625	0.0625
94	Ho true	680	160	0.0375	0.0375	0.0313
95	Ho true	700	160	0.0375	0.0500	0.0500
96	Ho true	720	160	0.0437	0.0688	0.0625
97	Ho true	740	154	0.0390	0.0390	0.0325
98	Ho true	760	150	0.0533	0.0400	0.0400
99	Ho true	780	150	0.0533	0.0467	0.0400
100	Ho true	800	150	0.0733	0.0533	0.0533
101	Ho true	820	150	0.0000	0.0000	0.0000
102	Ho true	840	150	0.0200	0.0200	0.0200
103	Ho true	860	150	0.0333	0.0467	0.0467
104	Ho true	880	150	0.0400	0.0533	0.0467
105	Ho true	900	150	0.0267	0.0333	0.0333
106	Ho true	920	150	0.0867	0.0933	0.0867
107	Ho true	940	150	0.0733	0.0533	0.0533
108	Ho true	960	150	0.0467	0.0467	0.0467
109	Ho true	980	150	0.0333	0.0333	0.0333
110	Ho true	1000	150	0.0467	0.0267	0.0200
111	Ho true	1020	150	0.0600	0.0533	0.0533
112	Ho true	1040	150	0.0533	0.0533	0.0533
113	Ho true	1060	150	0.0733	0.0533	0.0467
114	Ho true	1080	150	0.0667	0.0800	0.0733
115	Ho true	1100	150	0.0933	0.0867	0.0867
116	Ho true	1120	150	0.0400	0.0267	0.0267
117	Ho true	1140	150	0.0600	0.0733	0.0667
118	Ho true	1160	150	0.0267	0.0400	0.0400
119	Ho true	1180	150	0.0533	0.0600	0.0533
120	Ho true	1200	150	0.0667	0.0600	0.0533

Appendix E.

Pascal Program used to "Condense" SAS printouts of results

```
{U+}
PROGRAM Brutel8 ;
(* ***** *)
(* SRC_Name and OUT_Name designate input and output files. *)
(* Infile is the printout. Outfile is the filtered data. *)
(* ***** *)

CONST
  SRC_Name = 'brutel8.lst' ;
(* OUT_Name = 'con:' ; *)
  OUT_Name = 'brutel8.out' ;

TYPE
  LinoText = STRING[150] ;
  FinString = STRING[8] ;

VAR
  InputLine, OutWord, OutputLine : LinoText ;
  InFile, Outfile : Text ;
  i : integer ;
(* ***** *)
(* ***** *)
(* ***** *)

PROCEDURE Initialize ;

BEGIN
  Assign(Infile, SRC_Name) ;
  Assign(Outfile, OUT_Name) ;
  Reset(Infile) ;
  Rewrite(Outfile) ;
END ;
(* ***** *)

PROCEDURE getn ;

BEGIN
  IF (pos('Number of observations in data set', inputline) = 1) THEN
    BEGIN
      write('.');
      writeln(Outfile, 1:3, ' ', Inputline, ' ');
      FOR i := 1 TO 6 DO
        BEGIN
          readln(Infile, Inputline) ;
          CASE i OF
            4 : BEGIN
                  delete(Inputline, 1, 1) ;
                  delete(Inputline, 38, length(Inputline) - 38) ;
                  writeln(Outfile, 2:3, ' ', Inputline, ' ') ;
                END ;
          END ;
        END ;
      END ;
    END ;
  END ;
(* ***** *)
```

PROCEDURE getvars ;

```

BEGIN
  IF (pos('Dependent Variable',inputline) = 1) THEN
    BEGIN
      writeln(Outfile,3:3,'    ',Inputline,'    ');
    END ;
  END ;
(* ***** *)

```

PROCEDURE getman1 ;

```

BEGIN
  IF (pos('the Hypothesis of no WAVE*SITE',inputline) = 1) THEN
    BEGIN
      writeln(Outfile,4:3,'    ',Inputline,'    ');
      FOR i := 1 TO 9 DO
        BEGIN
          readln(Infile,Inputline) ;
          CASE i OF
            8 : BEGIN
                  delete (Inputline,23,17);
                  writeln(Outfile,5:3,'    ',Inputline,'    ');
                END ;
          END ;
        END ;
      END ;
    END ;
  END ;
(* ***** *)

```

PROCEDURE getws ;

```

BEGIN
  IF (pos('Source: WAVE*SITE',inputline) = 1) THEN
    BEGIN
      writeln(Outfile,6:3,'    ',Inputline,'    ');
      FOR i := 1 TO 13 DO
        BEGIN
          readln(Infile,Inputline) ;
          CASE i OF
            2 : BEGIN
                  delete (Inputline,1,45);
                  writeln(Outfile,7:3,'    ',Inputline,'    ');
                END ;
            3 : BEGIN
                  delete (Inputline,1,45);
                  writeln(Outfile,8:3,'    ',Inputline,'    ');
                END ;
            10 : BEGIN
                  writeln(Outfile,9:3,'    ',Inputline,'    ');
                END ;
            11 : BEGIN
                  writeln(Outfile,10:3,'    ',Inputline,'    ');
                END ;
          END ;
        END ;
      END ;
    END ;
  END ;
(* ***** *)
(* ***** main ***** *)

```

```
(* ***** m a i n ***** *)  
(* ***** *)
```

```
BEGIN
```

```
  Initialize ;
```

```
  WHILE NOT eof(infile) DO
```

```
    BEGIN
```

```
      readln(infile,inputline) ;
```

```
      getn ;
```

```
      getvars ;
```

```
      getman1 ;
```

```
      getws ;
```

```
    END ;
```

```
  writeln(Outfile) ;
```

```
  Close(Outfile)
```

```
END.
```